Query Routing in Cooperative Semi-structured Peer-to-Peer Information Retrieval Networks

Rami Suleiman Ayed Alkhawaldeh

School of Computing Science
College of Science and Engineering
University of Glasgow

This dissertation is submitted for the degree of
Doctor of Philosophy (PhD)

14^{th} October 2016

Dedication

To my dear father,
Who is playing a great role in my life, supporting, encouraging, and teaching me patience and challenge to face the difficulties and obstacles

To my merciful mother,
Who is still lightening my life as a candle by her warm spirit

To my wife and my children,
Who is never ending her moral support and prayers which always acted as a catalyst in my academic life

To my sisters and friends,
Who are still encouraging and inspiring me through my studying years

To the ones,
Whom I respect and never forget, who offer me so many great things

To all those,
I dedicate this work which I hope will have your satisfaction and attention.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work, under the supervision of Joemon M. Jose & Inah Omoronyia, and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. Permission to copy without fee all or part of this thesis is granted provided that the copies are not made or distributed for commercial purposes, and that the name of the author, the title of the thesis and date of submission are clearly visible on the copy.

Rami S. Alkhawaldeh
October, 2016

Acknowledgements

Undertaking this Ph.D. has been a truly life-changing experience for me, and it would not have been possible to do without the support and guidance that I received from many people.

First of all, I would like to express my special appreciation and thanks to my supervisor Professor Joemon M. Jose, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. I would also like to thank my second supervisor, Dr. Inah Omoronyia, for your feedback with respect to my yearly reports and vivas. I would also like to thank my committee members, Prof. Udo Kruschwitz and Dr. Frank Hopfgartner for their insightful comments and suggestions, but also for the hard question which incited me to widen my research from various perspectives.

I will forever be thankful to my research advisor, Dr. Deepak Padmanabhan. Deepak has been helpful in providing advice many times during my Ph.D. journey. Secondly, I would like to thank those in my office for making the experience so enjoyable: Philip, James, Jesus, Fajie, David Maxwell, Jorge Paule & Stewart.

I would also like to say a heartfelt thank you and sincere gratitude to my mum Mrs. Fatima Khawaldeh for her indefatigable patience, my Dad Mr. Suleiman Alkhawaldeh for always believing in me and encouraging me to follow my dreams, and my sisters Rania, Lana and Dana for their support and encouragements during my Ph.D. journey.

Finally, I would like to acknowledge the most important people in my life -my wife Samar and my children Hamzah and Mira. They have been a constant source of strength and inspiration. There were times during the past four years when everything seemed hopeless and I did not have any hope. I can honestly say that it was only their determination and constant encouragement.

Abstract

Conventional web search engines are centralised in that a single entity crawls and indexes the documents selected for future retrieval, and the relevance models used to determine which documents are relevant to a given user query. As a result, these search engines suffer from several technical drawbacks such as handling scale, timeliness and reliability, in addition to ethical concerns such as commercial manipulation and information censorship. Alleviating the need to rely entirely on a single entity, Peer-to-Peer (P2P) Information Retrieval (IR) has been proposed as a solution, as it distributes the functional components of a web search engine – from crawling and indexing documents, to query processing – across the network of users (or, peers) who use the search engine. This strategy for constructing an IR system poses several efficiency and effectiveness challenges which have been identified in past work. Accordingly, this thesis makes several contributions towards advancing the state of the art in P2P-IR effectiveness by improving the query processing and relevance scoring aspects of a P2P web search.

Federated search systems are a form of distributed information retrieval model that route the user's information need, formulated as a query, to distributed resources and merge the retrieved result lists into a final list. P2P-IR networks are one form of federated search in routing queries and merging result among participating peers. The query is propagated through disseminated nodes to hit the peers that are most likely to contain relevant documents, then the retrieved result lists are merged at different points along the path from the relevant peers to the query initializer (or namely, customer). However, query routing in P2P-IR networks is considered as one of the major challenges and critical part in P2P-IR networks; as the relevant peers might be lost in low-quality peer selection while executing the query routing, and inevitably lead to less effective retrieval results. This motivates this thesis to study and propose query routing techniques

to improve retrieval quality in such networks.

Cluster-based semi-structured P2P-IR networks exploit the cluster hypothesis to organise the peers into similar semantic clusters where each such semantic cluster is managed by super-peers. In this thesis, I construct three semi-structured P2P-IR models and examine their retrieval effectiveness. I also leverage the cluster centroids at the super-peer level as content representations gathered from cooperative peers to propose a query routing approach called Inverted PeerCluster Index (IPI) that simulates the conventional inverted index of the centralised corpus to organise the statistics of peers' terms. The results show a competitive retrieval quality in comparison to baseline approaches. Furthermore, I study the applicability of using the conventional Information Retrieval models as peer selection approaches where each peer can be considered as a big document of documents. The experimental evaluation shows comparative and significant results and explains that document retrieval methods are very effective for peer selection that brings back the analogy between documents and peers. Additionally, Learning to Rank (LtR) algorithms are exploited to build a learned classifier for peer ranking at the super-peer level. The experiments show significant results with state-of-the-art resource selection methods and competitive results to corresponding classification-based approaches.

Finally, I propose reputation-based query routing approaches that exploit the idea of providing feedback on a specific item in the social community networks and manage it for future decision-making. The system monitors users' behaviours when they click or download documents from the final ranked list as implicit feedback and mines the given information to build a reputation-based data structure. The data structure is used to score peers and then rank them for query routing. I conduct a set of experiments to cover various scenarios including noisy feedback information (i.e, providing positive feedback on non-relevant documents) to examine the robustness of reputation-based approaches. The empirical evaluation shows significant results in almost all measurement metrics with approximate improvement more than 56% compared to baseline approaches. Thus, based on the results, if one were to choose one technique, reputation-based approaches are clearly the natural choices which also can be deployed on any P2P network.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "I have not failed. I have just found 10,000 things that do not work."
>
> — Thomas Edison, (1847-1931)

1.1 Searching the Web

With the World Wide Web (or, *web*) playing an increasingly central role in our daily lives, it is growing exponentially with an overwhelming amount of multimedia content such as text, images, audio and video. One of the biggest challenges faced by users is Information Retrieval (IR), that is, finding pages containing information they are seeking. Accordingly, the vast amount of information must be constantly organised to be made accessible for the nearly 3.6 billion people who use the web during their daily lives[1].

Users have diverse information needs that range from navigation to a known resource (e.g., "`bbc news website`") to transactional (e.g., "`where can i buy a flight`") and informational (e.g., "`places to visit in london`") (Broder, 2002). A web search engine provides a starting point where a user can express their information need as a *query* comprising descriptive terms expected to be found in relevant items. Following query submission, a search engine automatically provides a ranked list of relevant search results for the user to examine. Results presented are drawn from the sample of content the search engine has previously discovered through a crawler that endlessly follows links from page-to-page across the web for new information. Descriptive terms in each web page are

[1] http://www.internetworldstats.com/stats.htm (October, 2016)

indexed, ready for matching with submitted user queries. Highly-linked content, termed the *surface web*, is easily discovered by the crawler and is thus readily available for discovery through conventional web search engines.

However, since the web is growing so quickly, it is infeasible for search engines to crawl all content, all the time (Rudesill et al., 2015). New content can take time to be discovered, and some content may never be found since it is hidden deep within vast websites such as databases, web forums, webmail pages, and pages behind paywalls, with few or no incoming links (He et al., 2013; Rudesill et al., 2015). Furthermore, as the web increasingly grows, it becomes impractical to maintain the vast computing and storage resources necessary to provide a complete index of the web (Rudesill et al., 2015). As a result, conventional search engines are unable to search much of the content available on the web (He et al., 2013). This un-searchable, yet potentially valuable, content is otherwise known as the *deep web*. Recent estimates suggest that the deep web contains 400-500 times more public information than the surface web (Bergman, 2001; Rudesill et al., 2015) – meaning a large amount of potentially valuable information is excluded from search results.

Beyond the issue of indexing web content ready for search, to achieve user satisfaction a search engine must also satisfy retrieval quality and query response time expectations. This poses serious competing challenges for search engine effectiveness and efficiency (Croft et al., 2009). The retrieval effectiveness measure estimates the ability of the system in providing the correct information that caters user's information needs (i.e, retrieval quality). The efficiency problem focuses on the response time of the system when to retrieve the result set (i.e, query response time). Effectiveness may be inhibited by the fact Internet users often find it difficult to express their information need. Because of this, they often provide ambiguous queries which are on average only 2.23 words in length (Nguyen et al., 2007; Fang et al., 2011). This makes it challenging to determine document relevance. Efficiency meanwhile is complicated by the fact web search engines comprise of large, complex systems numbering many thousands of servers interconnected through different networks and scattered into multiple data centres to handle large volumes of queries per second (Croft et al., 2009; Baeza-Yates and Ribeiro-Neto, 2011).

Conventional web search engines are built upon client-server architectures; with the centralised server maintained by the web search engine company coordinating all index management and query processing tasks. Servers store large amounts of information and have the ability to deliver or "serve" that information quickly and efficiently due to their high processing and storage capabilities, while the clients represent the end-users who benefit from the servers' services. IR systems perform many functions. In particular, they represent, store, and organise the documents in an accessible manner through a process called indexing, and then provide a query processing interface for the users. The indexing process starts extracting the words (or, tokens) from documents (Salton and McGill, 1986; Baeza-Yates and Ribeiro-Neto, 2011), eliminating the common tokens (or stop words) and then pruning them into their grammatical root (Porter, 1980; Peng et al., 2007). Finally, an analyser produces statistical information on the processed terms given their related documents into two data structure files for query processing (Luhn, 1957; Salton et al., 1975; Dean, 2009). From the client-side perspective, users request a search engine for specific information which is, in turn, retrieves the relevant documents for the given query satisfied their information need.

Although centralised search engines have advantages of simplicity in document management and high efficiency in comprehensive search, they are susceptible to ethical and technical drawbacks that are varied from scalability and user privacy risk to its weaknesses in crawling the deep web content, which is substantially larger than the indexable surface web (Bergman, 2001; Lewandowski et al., 2006; Tene, 2009). It would be better if users and creators of web content could collectively provide a search service and have full control over what information they wish to share as well as how they share it. Thereby, in order to address the search engines' shortcomings, the scope of the thesis is to investigate the searching process using federated search systems (distributed information systems), especially P2P networks that are touted as an alternative framework.

1.2 Federated Search and P2P-IR Networks

Federated (or similarly, *distributed*) search engines have emerged as a promising paradigm to alleviate the aforementioned drawbacks. These systems provide a uniform interface across a plurality of searchable resources by way of a broker (Shokouhi and Si, 2011). The broker submits a query in parallel to these resources (or text collections) that have a high probability of relevant documents. The retrieved result lists of the selected collections are then merged into a final result list for users to cater their information needs. There are three forms of federated search systems (Shokouhi and Si, 2011): meta-search, vertical (or aggregated) search, and Peer to Peer (P2P) network search. In meta-search, the broker sends a given query in parallel to multiple *search engines* and combines the retrieved result lists into a final ranked list (Meng et al., 2002). In vertical search, the broker sends a query to a set of search verticals (e.g, images, news, blogs, books, videos, and maps) often different in topics and incorporates the retrieved multimedia answers along with the default text results into a final ranked list (Hawking, 2004; Bailey et al., 2007; Kopliku et al., 2014). The P2P network search is considered one of the federated search systems in sending a query to multiple resources (or peers) and merging the retrieved result lists along the path from the responding resources to the query sender (or customer) (Tigelaar et al., 2012; Klampanos and Jose, 2012).

Peer to Peer overlay networks presume that the users on the web play the role of a client and server at the same time to store content and request for information. The architecture of P2P networks is built logically over physically connected nodes located at the edge of the Internet. However, P2P Information Retrieval (P2P-IR) is one type of federated search; as the approach has a number of similarities with that of distributed information retrieval systems (Lu, 2007; Klampanos and Jose, 2012; Tigelaar et al., 2012). There are three major challenges in P2P-IR systems that have to be taken into account to design an efficient and effective retrieval system, which are resource representation, resource selection (or query routing) and result merging (or fusing). Resource representation refers to the process of acquiring resource description that is used for resource selection (or query routing). Query routing is the process of ranking peers based

on their representations to a given query and sending the query to most relevant ones. The requested peers response with a set of result lists that are merged by peers (or super-peers) into a final ranked list in an unknown process called result merging (or fusing). This thesis focuses on query routing in a specific type of P2P-IR networks. Query routing is a critical component in P2P-IR; low-quality resource selection, the case where the relevant peers get excluded would inevitably lead to less effective IR results. One of the difficulties in P2P architectures is that it is almost impossible to collect the global statistics, which are needed to be estimated to route a query to the relevant peers (Richardson and Cox, 2014). Hence the lack of global statistical information leads to flooding the P2P-IR networks with queries that result in high computation costs to process these queries, high bandwidth limits, and increases in non-relevant documents in the final merged result list. Consequently, reducing the number of messages adversely affect the effectiveness of the system. Therefore, in order to build effective P2P retrieval systems and tackle such poor retrieval performance, alternative mechanisms are needed.

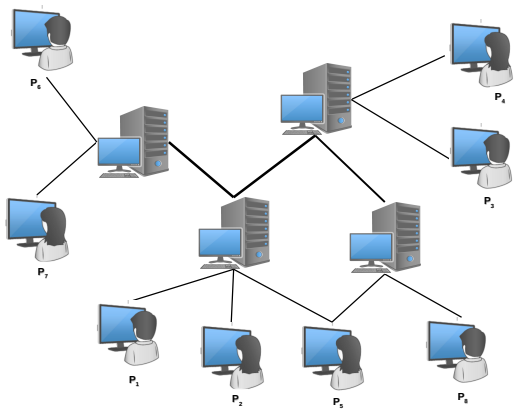


Figure 1.1: Semi-structured P2P Architecture.

The cluster hypothesis postulates that grouping similar documents into semantic groups leads to efficient retrieval results (van Rijsbergen, 1979). An effective and efficient P2P-IR network exploits the clustering hypothesis to huddle the peers coherently with similar domain interest around the same cluster (Klampanos and Jose, 2007; Lu, 2007). Consequently, the distance between similar peers

is short where a query is routed via a shortest path as in small-world networks (Watts and Strogatz, 1998; Kleinberg, 2000). In such scenarios, it is worth to use the clustering algorithms to form a small-world network to increase the search efficiency and reduce the message routing overhead (Lu, 2007; Klampanos and Jose, 2003). The semi-structured P2P overlay network is a cluster-based topology that exploits the heterogeneity of nodes with regard to their robustness and capacity to fairly distribute labour on the system. This network is proposed as a promising structure to build retrieval approaches, which contains two types of peers; super (or hub) and regular peers (users) (Klampanos and Jose, 2007; Lu and Callan, 2003) as shown in Figure 1.1. The super-peers have a high level of willingness to store the meta-data of their associated regular peers and communicate with each other to cast queries on behalf of their own regular peers (Tigelaar et al., 2012). Hence, the semi-structured P2P network combines the advantages of the two centralised and decentralised P2P overlay networks in load balancing between super and regular peers and through providing heterogeneity across peers to improve the performance (Klampanos and Jose, 2007; Tigelaar et al., 2012).

Semi-structured P2P-IR networks provide a coherent system to effectively and efficiently route a query and mitigate the poor retrieval performance in general P2P networks. The assumption, given the semi-structured P2P-IR networks, is that the retrieval effectiveness and query routing efficiency can be improved by exploiting such coherent clusters between the peers, using resource selection methods and/or proposing an efficient and effective peer ranking system to traffic a query to the relevant and highly ranked peers.

1.3 Problem statement

This thesis tackles the problem of low-quality retrieval performance in unstructured, cluster-based P2P-IR networks. In general, previous research has strived to mitigate the poor performance of P2P networks using a set of techniques varying from interest-based, content-based and small-world networks as topological architectures to enhancing the peer representation, query routing and merging techniques (Tigelaar et al., 2012; Klampanos and Jose, 2012). In contrast to this

previous work, in this thesis I examine cluster-based semi-structured P2P-IR networks that exploit the content-based approaches to group the peers into similar semantic domains using clustering techniques in order to alleviate the challenge of low retrieval quality (Klampanos and Jose, 2007; Lu, 2007).

However, the robustness of semi-structured P2P networks resides in combining the peers into coherent groups, whereas a query could be routed within homogeneous peers of the same topic. This topological architecture has an advantage of using highly robust and willing super-peers to work on behalf of their peers to route query and retrieve relevant documents. In addition, the churn rate of peers, when joining and leaving the systems, has little impact on the retrieval quality as there are more relevant documents under super-peer level within the same semantic peers (Alkhawaldeh and Jose, 2015). Although these topologies group similar documents within the peers at the super-peer level, they suffer from routing a query to the most likely relevant peers. The reason is the noise in clustering web documents, which results in a weak representation of clustered peers (Klampanos and Jose, 2007).

Due to the importance and the advantages of deploying semi-structured P2P-IR networks for Information Retrieval (Klampanos and Jose, 2007; Alkhawaldeh and Jose, 2015), the motivation is to explore solutions to improve the retrieval effectiveness using these networks through examining a set of factors that might have an effect in this deficiency of the retrieval quality. Besides these factors, I also examine a set of resource selection methods on these networks as effective solutions. Finally, I exploit the user behaviour in real-life P2P networks in providing feedback during the interaction with the system for enhancing the retrieval quality. In summary, the objective is to improve the performance of semi-structured P2P-IR networks in different parameter settings, resource selection methods, and user behaviour scenarios.

1.4 Challenges

In general, studying the query routing in semi-structured P2P networks highlight several challenges. This section outlines a set of challenges that have to be taken into consideration in designing an effective query routing in semi-structured P2P-

IR networks, which are:

- The dynamic nature of P2P networks has an effect on aggregating the global term statistics for estimating the relevance score of peers (Richardson and Cox, 2014). The peers have the ability to join and leave the system at any time, which has an impact on the distribution of documents and leads to low-quality of routing a query in the system (Stutzbach and Rejaie, 2006; Tian and Dai, 2007). Hence, there is a need for an approach to find these global statistics to effectively and efficiently traffic a query to most likely relevant peers, which is a challenging problem.

- An efficient and effective P2P network routes a query in shortest paths to the relevant peers that are most promising to evaluate it. There have been several architectures proposed for organising the peers to achieve such a goal. Cluster-based architecture is considered as one of the promising topologies in full-text P2P-IR networks (Xu and Croft, 1999; Lu, 2007; Klampanos and Jose, 2007). However, using clustering techniques in P2P networks to group the peers into similar semantic domain is a challenging problem. Using an ineffective clustering algorithm might lead to poor grouping of peers and further adversely affect routing a query to relevant peers. In addition, even though each peer might have a small number of documents and with its own computation power, the exponential growth of the web increases the complexity in clustering this huge amount of content.

- In P2P networks, no standard testbeds and metrics exist for P2P-IR evaluation. Klampanos et al. (2005) proposed a number of P2P-IR testbeds based on real-life scenarios. These testbeds simulate three environments of P2P networks, which are file-sharing, digital library, and uniformly distributed environments. Although these testbeds have a set of properties such as each peer shares a small number of topics, distribution of documents in the system as a power-low pattern, and replication some contents across peers, the evaluation is still a neglected task and challenging issue in P2P-IR systems.

- In this thesis, I study the query routing in cooperative P2P-IR environments. In the cooperative environments, each peer should provide a statisti-

cal lexical information about its documents. Each peer could be represented as a big document of local documents with different topics. The challenge is how can we deploy the conventional retrieval models as resource selection methods for query routing in such environments. In addition, under this scenario, how can we exploit the Learning to Rank (LtR) algorithms to enhance the query routing quality (Liu, 2011).

- One of the main parts of this thesis is to estimate the reputation value for each peer based on the rated and retrieved relevant documents and then route a query based on these values. The reputation concept is used for security in P2P networks to prevent malicious behaviours of peers from harming and destroying the system or even encouraging selfish peers to provide and share their contents (Jøsang et al., 2007). In addition, in security aspect, the trust concept is used to retrieve relevant and trustworthy documents through using users' feedback in the ranked results list (Zhang, 2011). This demonstrates reputation and/or trust as psychological concepts which can be used in different manners as a selection criterion. The challenge is to use the reputation values to enhance query routing, which in turn improves retrieval efficiency and effectiveness in semi-structured P2P-IR networks. Accordingly, the sub-challenges are how to monitor the users' behaviours in providing feedback in the systems and how to mine and organise this feedback as a reputation-based data structure for future queries. Furthermore, it is not clear the effect of non-relevant documents on the concept of reputation, which in turn might have an effect on the reputation values of the peers and biases the routing of a query to non-relevant peers. Finally, an important challenge is to study the behaviours of users in abstaining from providing feedback on the retrieved result list as lacking usage information for building the reputation. Different scenarios should be studied to cover these challenges and to show the importance of using the reputation concept to enhance the performance of semi-structured P2P-IR networks.

1.5 Structure and Contribution of the Thesis

This thesis makes a set of contributions to improve the query routing techniques in semi-structured P2P-IR networks. This section discusses the remainder chapters of the thesis along with a research question in each contribution chapter. These research questions are presented as high-level (HL). The sub-research questions are presented in corresponding contribution chapters.

- **Chapter 2: Information Retrieval systems** provides a comprehensive review of web search engines, web search ranking, information retrieval evaluation, and the advantages and disadvantage of search engines. In particular, I explain the process of extracting documents and organising them in data structures for query processing in the future. A set of weighting and ranking algorithms are clarified, including boolean, probabilistic, language model, and LtR algorithms that assign a score to the documents that are most likely relevant to the given query. In addition, I discuss evaluation methodologies to evaluate information retrieval systems to enhance their components. I also discuss relevance feedback information. Finally, I exhibit the advantages and disadvantages of search engines as a motivation to explore the objectives to improve the performance of web search.

- **Chapter 3: Distributed Information Retrieval** presents the federated search concept as a distributed information retrieval system and explains the environments, the forms, and the challenges in this distributed search paradigm. The chapter also expands the discussion of the importance of using P2P networks for information retrieval. In particular, I discuss a set of cooperative resource selection methods used for federated search, especially in meta-search environments, as well as the state-of-the-art classification-based resource selection methods. In addition and in more detail, I discuss the P2P network, including its concept, architectures, topologies, and more specifically the information retrieval task in these networks. Finally, the advantages and disadvantages of P2P-IR networks are followed.

- **Chapter 4: Semi-structured Peer to Peer Information Retrieval** explains the dataset varies from document representation, testbeds, query

set, and evaluation metrics to experimental settings and parameters. Due to the importance of studying cluster-based semi-structured P2P-IR networks, the chapter demonstrates the process of building the semi-structured P2P-IR network through two steps of clustering, which are intra-peer and inter-peer clustering. In intra-peer clustering, the documents in each peer are grouped into different semantic clusters and the computation cost of doing clustering is studied. Inter-peer clustering utilises the in-peer centroids to another level of clustering to combine the peers into semantic groups. In inter-peer clustering, I present the process of building three different topological architecture models of the semi-structured P2P-IR network at the super-peer level and study the performance boundaries and the effectiveness of their information retrieval. After discussing the three models, I select the most competitive and effective model and use it to examine different design considerations on retrieval quality such as conventional retrieval models that are applied in each peer, the number of super-peers, and the churn rate of the peers in joining and leaving the system (or failure and departure of peers in the system).

HL-RQ1: *How does the effectiveness change based on the way in which semi-structured P2P-IR networks are constructed, and how do the parameters in forming such networks influence the effectiveness?*

- **Chapter 5: Cooperative Resource Selection Methods in Federated Search** represents the retrieval effectiveness and routing efficiency results of using cooperative resource selection methods on the effective semi-structured P2P-IR model as I studied in Chapter 4. In particular, I exploit the coherent lexical cluster centroids at the super-peer level to build an inverted index for query routing in such specific networks which is called Inverted PeerCluster Index (IPI). Moreover, I utilise the concept of big documents and cluster hypothesis to examine the applicability of conventional IR models as resource selection methods. Hence I do an empirical benchmark of document retrieval methods to inspect their effectiveness and efficiency under the coherent semantic resources and meta-search environments. Furthermore, I investigate the importance of using the state-of-the-

art Learning to Rank (LtR) algorithms as resource selection approaches in meta-search environment and semi-structured P2P-IR networks.

HL-RQ2: *How can we build effective query routing algorithms in semi-structured P2P-IR scenarios?*

- **Chapter 6: Reputation-based Query Routing** discusses reputation as a social concept and expands the definition to reputation relevance of objects that might be documents and/or peers. In prior research work, the users' behaviours in P2P networks was used to enhance the security perspective through punishing malicious and selfish peers in the system (Jøsang et al., 2007). The users provide feedback on a specific object and the system aggregates this feedback from trusted peers as reputation values for future interaction (Abdul-Rahman and Hailes, 2000). In this chapter, through exploiting the reputation relevance concept, I particularly explain the process of building a reputation-based data structure to be used as a query routing for future queries. The reputation-based data structure that is stored and organised by a super-peer is used to propose a set of resource selection methods that are naturally and most likely to be used as query routing in the semi-structured P2P-IR networks. The reputation-based data structure depends on implicit users' feedback (i.e, click through data) from past interactions in the system, which is simulated using a training query set. In evaluating reputation-based query routing approaches, I examine five scenarios of evaluating methodologies, in addition to the retrieval efficiency. The four scenarios that are simulated for retrieval effectiveness include: (i) simulating the methods on all usage reputation information given by the users, (ii) using lack of usage information by excluding the reputation information of testing queries, (iii) expanding the leave-out usage information, providing noisy feedback from the users, and (iv) combining the reputation-based methods with other resource selection methods.

HL-RQ3: *Is implicit feedback provided by the users during their interactions in semi-structured P2P-IR networks effective for improving query routing, and how should such feedback be exploited to build reputation*

data structures?

- **Chapter 7: Conclusions and Further Work** reviews the thesis contents, contributions, and potential future works. In particular, I discuss the conclusions of each chapter and open a motivation for the next chapter. I also discuss the future work along with future research questions.

1.6 Publications

The research presented in this thesis is contained in several first-author publications. These are as follows:

1. **Rami S. Alkhawaldeh** and Joemon M. Jose. Experimental study on semi-structured peer-to-peer information retrieval network. In CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, pages 3–14. **(Full paper)** Chapter 4.

2. **Rami S. Alkhawaldeh**, Joemon M. Jose, and Deepak P. Clustering-based Query Routing in Cooperative Semi-structured Peer to Peer Networks. In ICTAI 2016, San Jose, USA November 06-08, 2016. **(Short paper)**, Chapter 5.

3. **Rami S. Alkhawaldeh**, Joemon M. Jose, and Deepak P. Evaluating Document Retrieval Methods for Resource Selection in Clustered P2P IR. In CIKM 2016, Indianapolis, USA October 24-28, 2016, Proceedings, pages 2073–2076. **(Short paper)** Chapter 5.

4. **Rami S. Alkhawaldeh**, Joemon M. Jose, Deepak P., and Fajie Yuan LTRo: A Learning to Route Approach in Cooperative Semi-structured P2P networks. In ECIR 2017, Aberdeen, Scotland UK April 8-13, 2017 **(Short paper)** -Accepted, Chapter 5.

-Others:

5. Fajie Yuan, Joemon M. Jose, Guibing Guo, Long Chen, Haitao Yu, **Rami S. Alkhawaldeh**. Joint Geo-Spatial Preference and Pairwise Ranking for Point-of-Interest Recommendation. In ICTAI 2016, San Jose, USA November 06-08. **(Full paper)**.

Part I
Background and Literature Review

This thesis focuses on Information Retrieval on semi-structured P2P networks. In the background and literature Review, I survey background information about Information Retrieval systems, including the process of managing document and query processing to retrieve relevant documents for a given query. In addition, I explain the evaluation methodologies for evaluating the retrieval effectiveness of retrieval systems and also clarify the relevance information feedback followed by the advantages and disadvantages of using search engines. In the second chapter of this part, I present the P2P networks concept, architectures, topologies, Reputation-based systems, and focus on information retrieval process in such networks through providing related works.

Chapter 2

Information Retrieval Systems

"Much learning does not teach understanding."
—Heraclitus, (544-483 B.C)

2.1 Introduction

Over the last few decades, Information Retrieval (IR) systems have emerged as a system of managing information items through representing, storing, and organising them in an accessible manner (Baeza-Yates and Ribeiro-Neto, 2011). The goal of IR systems is to provide items that are relevant to users' information needs. The items, in this thesis, are structured or semi-structured documents while the information need is formulated as a natural language query. However, the main challenge in IR systems is to clarify the relevance of documents to a given user's query (Goffman, 1964). Relevance is a complex concept that can be defined in two aspects; which are topical and user relevance as well as binary and multivalued relevance (Manning et al., 2008a). In topical relevance, the document is judged to be relevant if it matches a given query topic while user relevance assesses the documents based on other factors that support the document relevance such as the date, language, and author of the document. Binary and multivalued relevance determine how much is the document relevant to the user's query. In particular, binary relevance rates the documents to be relevant or non-relevant whereas multivalued relevance assesses the documents based on a varied range of relevance, such as most relevant, relevant, less relevant, non-relevant, and unsure.

Several efforts have been exerted to achieve the goal of determining the relevance of a document's text to the user's information need formulated as a query of terms. The theory of relevance can be represented as mathematical retrieval models that match the given query to documents and rank them as a list of results. The effective retrieval model is the one that ranks the relevant documents at the top of the retrieved, ranked list. The retrieval effectiveness of retrieval models depends on the ability of its ranking algorithm to retrieve a high number of relevant documents.

Web search engines are considered as one of the crucial applications of IR systems that have a set of challenges include massive-scale of documents, heterogeneity of the produced content, and interconnected nature of the web (Croft et al., 2009). In order to tackle these challenges, web search engines are essentially composed of three core elements: crawler, indexer, and query processor as depicted in Figure 2.1. In particular, the crawler finds and extracts documents into a centralised corpus (or collection). The indexer creates efficient data structures (indices) from the collection to facilitate the content access. The query processor uses the data structures to produce a ranked list of documents that are relevant to the user's query. These components along with web search ranking and evaluation will be discussed in more details in this chapter.

The chapter is organised as follows: Section 2.2 discusses web search engines' components, Section 2.3 discusses web search ranking algorithms, Section 2.4 discusses the evaluation metrics in IR systems. The relevance information feedback is dicussed in Section 2.5. Finally, the advantages and disadvantages of centralised search engines will be discussed in Section 2.6.

2.2 Web Search Engines

The search engine components work together in a cooperative manner to meet a user's information need. These components crawl the web to prepare a corpus of documents, build a data structure to organise these documents and then provide a query processing interface to facilitate the searching process for the users. This process depends on the performance of search engine components to achieve such goal. The remainder of this section describes briefly the three components as

depicted in Figure 2.1.

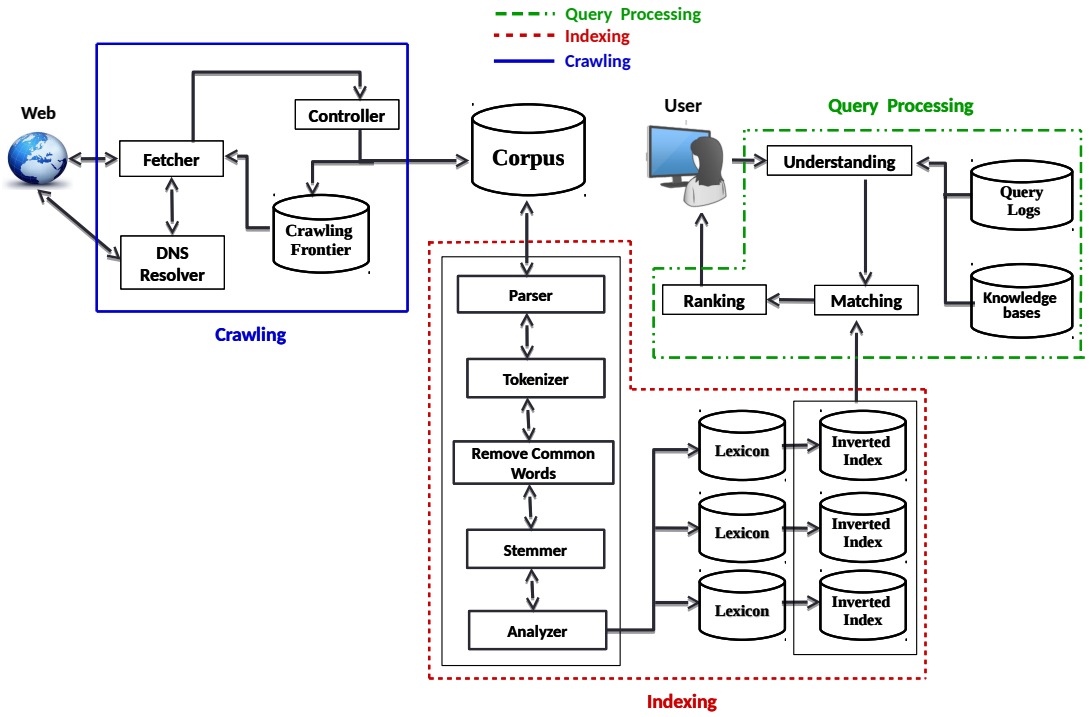


Figure 2.1: Search engine architecture

2.2.1 Crawling

Crawling is a process of collecting and finding documents on the web and store them in a local corpus for indexing. The crawler has a responsibility for fetching the documents from the web, which aims to feed the corpus with a maximum rate of documents in a short possible time (Pant et al., 2004; Castillo, 2004). In order to achieve that, the crawlers in search engines comprise of a set of components that cooperate with each other in a traversal algorithm to retrieve a set of documents for indexing (Castillo, 2004); which are crawling frontier, DNS (Domain Name Service) resolver, fetcher, and controller. The *crawling frontier* is initially filled with a list of URL seeds to be visited. The *DNS* resolver translates the URL domain into an IP address. The *fetcher* extracts next URL from the frontier and download it by using its IP address from the DNS resolver. The *controller*

processes the fetched documents where the extracted contents are stored locally for indexing in the corpus. The URLs that are extracted from the documents, in addition to their own URLs are inserted back into the frontier to be visited later for continuous crawling. Recently, however, the crawlers limit the search area on the web to find the relevant documents of a pre-defined set of topics using outstanding documents. This avoids irrelevant regions when fetching the documents for indexing, which is called focused crawling (Chakrabarti et al., 1999; Anagnostopoulos and Avraam, 2011). The focused (or topical) crawling requires less time, effort and cost processing to fetch the web documents that have undesirable value (Achsan and Wibowo, 2014). Although there is an efficient in extracting pre-defined relevant documents, the focused crawlers have still suffered from fetching hidden web (or unreachable web) (Achsan and Wibowo, 2014).

One of the challenges in crawling the web is gathering the hidden web (or deep web) (Bergman, 2001; Rudesill et al., 2015). The deep web is the content that is not reachable (or does not have a link to be fetched) by the crawlers. In contrast, the surface web is the reachable documents in the web graph. Deep web[1] is larger than the surface web, although the surface web is massive by itself, which increases the complexity for crawlers to extract these documents (Rudesill et al., 2015).

2.2.2 Indexing

The corpus contains a plain text of each document that is not suitable for technically matching the users' query terms. Thus, search engines use an indexer that converts the plain text of documents into appropriate data structures to be efficiently accessed by the users (Manning et al., 2008a; Croft et al., 2009). In particular, the indexer comprises a set of components that process and store the documents into organised indices, which include the parser, the tokenizer, the stop words remover, the stemmer, and the analyser as depicted in Figure 2.1. The parser extracts the content of each document into raw text. The document's text is then given to the tokenizer that splits it into individual tokens (Salton and McGill, 1986). The documents, in general, contain a set of tokens that have little

[1] https://hewilson.wordpress.com/what-is-the-deep-web/statistics/ (October, 2016)

discriminative power in identifying the relevance of a document to a given query such as stop-words (e.g. articles and connectives). Thus, search engines use the stop-words remover to discard these tokens from indexation to increase the retrieval efficiency and to save the storage cost. Consecutively, the tokens are then passed to another common operation called stemming (Porter, 1980; Peng et al., 2007). Stemming is a process of pruning the tokens into common grammatical root using a stemmer function. The stemmer function eliminates the affixes of a token that carry grammatical or lexical information to increase the probability of retrieving the documents that contain a variety of query terms and to reduce the size of indexing files (Baeza-Yates and Ribeiro-Neto, 2011). For instance, the terms "stemming", "stemmer", "stemmed" are reduced to their morphological roots "stem". In the final step, the analyser uses multiple text operations to generate the statistical information of these tokens such as term frequency and term document frequency (Luhn, 1957; Salton et al., 1975) and then builds two data structures to facilitate the query processing task. The two data structures are the lexicon file and inverted index (or sometimes inverted file) (Dean, 2009). The lexicon file stores information for all the unique terms in the corpus such as frequencies of terms along with the documents they occur. On the other hand, the inverted index stores for each term in the lexicon a posting list that contains information on the location of the term in different documents.

2.2.3 Query Processing

Query processing uses the indexed documents' statistics in order to match the query terms and then retrieves the most relevant documents using retrieval models to meet the user's information need (Arasu et al., 2001). The effective retrieval model is the one that ranks the retrieved documents and orders the most relevant ones on the top of the result list. Query processing includes three basic operations, which are query understanding, query matching, and document ranking. The query understanding process is essential for refining a query to reduce its poor representation to be as close as the user's information need (Li, 2010). Consequently, the query understanding component plays a major role in IR systems as low-quality query representation might result in deviating the retrieved content from the users' demands (Li et al., 2006; Kumaran and Carvalho, 2009). There

are different forms of query understanding operations that have been used for this purpose, such as query stemming (Porter, 1980; Peng et al., 2007), spelling correction (Ahmad and Kondrak, 2005; Li et al., 2006), terms deletion (Kumaran and Carvalho, 2009), name-entity recognition (Guo et al., 2009). Other useful operations include query topical classification to limit the number of retrieved documents (Shen et al., 2006) and query expansion that enhances a query representation by augmenting the query with useful terms from the local corpus or from outstanding resources like query log or knowledge base such as Wikipedia, WordNet, etc (Rocchio, 1971; Carpineto and Romano, 2012). Given the refined query terms, the query matching process retrieves the indexed documents that contain the query terms from the centralised indices. Finally, the document ranking process assigns scores to the retrieved documents and sets the most relevant ones to the top of the result list. In particular, the IR systems use mathematical retrieval models to determine the relevance of a document to the refined query terms. I will discuss in more details in the next section the ranking algorithms and their theoretical concepts in determining the relevance to documents.

2.3 Web Search Ranking

The user's query represented by fewer terms radically matches a large number of documents from a huge amount of documents in the web (Jansen et al., 2000b). These documents exceed the expected number of documents that have to be at the top of the result list. Therefore, the need for ranking algorithm is inevitably important to rank the matched documents and sets the most relevant ones on the top of the result list (Silverstein et al., 1999).

The document ranking process uses mathematical ranking functions to score the matched documents in descending order based on their relevance to the given query. Each ranking function has its own theory and signal in determining the relevance of a document. Hence there are three categories that score the relevance of the document to a query, which are query-dependent, query-independent, and query features. Query-dependent ranks the documents based on their signals to the query terms. Query-independent looks at other features that are far from the query itself, such as page rank (Brin and Page, 1998), content quality (Bender-

sky et al., 2011), spam likelihood (Cormack et al., 2011), URL Length, type etc. Lastly, query features take into consideration the query features solely to rank all the documents for an individual query such as the query topic classification, the history of the query in a query log, the predicted performance of the query, and the presence of entities such as persons and organisations in the query (Macdonald et al., 2012). In this section, I will focus on query-dependent, ranking algorithms. There are several query-dependent retrieval scoring models such as Boolean model, vector-space model and probabilistic model as I will describe in more details in the next subsections.

2.3.1 Boolean Retrieval

The Boolean retrieval model retrieves the documents that exactly match the specification of a query. This exact-matching retrieval model considers the relevance of documents as a binary decision whether the query specification is satisfied or not. Thus, this model assumes that all matched documents are the same in relevance as the query evaluation occurs in two possible values (TRUE and FALSE) and are determined by the logical Boolean operators (Manning et al., 2008a). Boolean models are used in several applications such as patent search (Joho et al., 2010) and legal search (Zhao and Callan, 2012) due to efficiency considerations (Kim and Croft, 2014). Since the Boolean models have a set of advantages such as the results are readable to the users, the query specification can include any data type such as document date and document type, and its efficiency in discarding the documents from the scoring process (Kim and Croft, 2014). In spite of these advantages, the Boolean retrieval models have a major drawback is that the retrieval effectiveness mainly depends on the users in formulating the given query (Croft et al., 2009). In particular, a simple query from the user retrieves all matched documents with little relevance, even the complex query, that is used to narrow the search result and retrieve the relevant documents, needs a reasonable experience from the user.

The Boolean retrieval models are sometimes used to retrieve a set of documents matched the given query because it is efficient and does not take more time to order these documents as the other ranking models (Li and Xu, 2012). Then specific approaches can be used to produce ranking scores for the documents

likely to be relevant to the user information need.

2.3.2 Term Weighting

The relevance of a document in the exact-matching model depends on the presence of a query in the document as a binary decision. Although Boolean models are important in some areas (e.g, patent and legal search), the Boolean models are inefficient in their retrieval ranking as two documents in the corpus contain the same query terms have the same relevance values. This means that the document with a high frequent number of query terms, that is likely to be more relevant, has the same relevance value to the document with a lower number of query terms (Luhn, 1957). Consequently, estimating the relevance is the problem of counting the frequency of query terms in the documents to be as indication weights to these documents (Salton and Buckley, 1988).

There are fundamental weighting and quantity schemes that represent the main core of query-dependent approaches. The schemes assign scores to the documents given a query and rank them based on these quantity scores. The quantity amount of a document is based on its relevance signal to the query terms such as term frequency ($tf_{t,d}$), document frequency ($df_{t,c}$), inverse-document frequency ($idf_{t,c}$). The term frequency ($tf_{t,d}$) determines the importance of the query term t to the document d through the number of occurrences of the term in the document. The document frequency ($df_{t,c}$) denotes the number of documents that contain the term t in the corpus c. Intuitively, the ranking algorithm assigns a high score to rare query terms that appear in few documents which are most likely to be relevant and scales down the common query terms occur too often in the corpus (Spärck Jones, 1972). The inverse-document frequency $idf_{t,c}$ quantity is defined to determine the nature of a term and reflects such phenomenon. Hence $idf_{t,c}$ denotes the power of the term t to discriminate the relevance between the documents in the corpus c. Formally, given the number of documents in the corpus N and the document frequency of the query term t (i.e, $df_{t,c}$), the inverse-document frequency is defined as:

$$idf_{t,c} = \log \frac{N}{df_{t,c}} \tag{2.1}$$

The idf weights of rare query terms obtain a high score, while the idf of common

query terms is likely to be low.

A composite strong weight can be defined through combining the two quantities term frequency and inverse-document frequency into one weight known as $tf_{t,d} \cdot idf_{t,c}$ weighting score. The $tf_{t,d} \cdot idf_{t,c}$ quantity assigns the term t in a document d with the highest score if the term t occurs many times in the documents with high discriminative power (Salton et al., 1975), which is defined as:

$$f(t,d) = tf_{t,d} \cdot idf_{t,c} \quad (or = tf_{t,d} \times idf_{t,c}) \tag{2.2}$$

In order to score all the query terms, a ranking algorithm can be defined as a scoring function using the previous quantities as follows:

$$Score(q,d) = \sum_{t \in q} f(t,d) \tag{2.3}$$

where $f(t,d)$ can be one of the weighting schemes that are discussed before.

Another commonly used quantity is document length (dl), which represents the number of tokens (regardless the frequency of term in the document) in document d as follows:

$$dl = \sum_{t \in d} tf(t,d) \tag{2.4}$$

The term weighting schemes are considered as the main core of query-dependent ranking approaches that are categorised as algebraic and probabilistic models (Baeza-Yates and Ribeiro-Neto, 2011) as I will discuss in more detail in the next subsections.

2.3.3 Vector Space Model

The vector space model was essentially the dominant retrieval model between the 60s and 70s, which is still used in these days. Documents, in the local corpus, are considered as a bag of words which can be represented as a vector of weights from the unique terms in the corpus. In the vector space model, several models can be used to calculate the similarity between two vectors of documents or document and query (Salton et al., 1975). The similarity can be estimated using the magnitude of the vector difference between two vectors. The drawback of this model occurs in the different length of two vectors as the two vectors have the same query content, this measure prefers the larger document. To tackle the effect of document length, we can use the cosine similarity between two vectors. For instance, given a document d and query q as vectors in the space of all unique

terms $t_i \in V$, such that:

$$q = (w_{t_{1,q}}, w_{t_2,q}, \cdots, w_{t_{v,q}}) \qquad and \qquad d = (w_{t_{1,d}}, w_{t_2,d}, \cdots, w_{t_{v,d}}) \tag{2.5}$$

where $v = |V|$ refers to the number of unique terms in two vectors, $w_{t,\bullet}$ denotes the weight of a term which might be $tf_{t,d}$, $idf_{t,c}$, or $tf_{t,d} \cdot idf_{t,c}$ for both vectors. A standard way to estimate the similarity is to compute the cosine similarity between the two vectors as follows:

$$Score_{VSM}(q,d) = cosine(q,d) = \frac{\vec{q} \cdot \vec{d}}{|q| \cdot |d|} = \frac{\sum_{i=1}^{v} w_{t_{i,q}} \cdot w_{t_{i,d}}}{\sqrt{\sum_{i=1}^{v} w_{t_{i,q}}^2} \cdot \sqrt{\sum_{i=1}^{v} w_{t_{i,d}}^2}} \tag{2.6}$$

where the numerator denotes the inner dot product of the two vectors whereas the denominator is the product of their Euclidean length as normalisation values. The value of $Score_{VSM}(q,d)$ is the cosine similarity value of the angle between two vectors in the vector space model. If the $Score_{VSM}(q,d)$ value is high, this means that the two vectors are close to each other and the angle is small, and vice versa. Therefore the system ranks the documents and retrieves more close documents to the query vector of high similarity. An un-normalised version of Equation 2.6 has been proposed with binary weights $w_{t,\bullet}$ as a simple, effective Coordination Level Matching (CLM):

$$Score_{CLM}(q,d) = \vec{q} \cdot \vec{d} = \sum_{i=1}^{v} w_{t_{i,q}} \cdot w_{t_{i,d}} \tag{2.7}$$

However, the probabilistic approaches use the same term weighting model to detect the relevance of a document to a query. These approaches exploit the probability theory to model a relationship between queries and documents as relevance to user's information need as I will describe in the next subsection.

2.3.4 Probabilistic Model

The relevance of a document for a given query is determined by the binary decision of the existence of the query terms in the document using the Boolean model while the relevance of the vector space model is determined by the similarity weight of the document and the query vectors such as cosine similarity in Equation 2.6. However, the relevance can be precisely identified as uncertainty quantity value of the whether the document has a content relevant to user information need.

The Boolean and vector space models depend on the indexed terms to compute the relevance, which is not sufficient to formulate the relevance uncertainty of the query to relevant and non-relevant documents.

The uncertainty inference of document relevance is an indication of how much the document contains relevant content based on the distribution of query terms in the relevant and non-relevant documents. The probability theory can be used as a robust principle to determine the uncertainty of occurring events which are, in IR perspective, the query terms in the documents. Hence the probabilistic retrieval models as a part of the retrieval process capture and evaluate the uncertainty of a document to a given query. The most dominated ranking principle is called the Probability Ranking Principle (PRP) (Robertson, 1977; Cooper, 1971), which is stated as:

> *"If a reference retrieval system's response to each request is* ***ranking of the documents*** *in the collections in* ***order of decreasing probability of relevance*** *to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis whatever data have been made available of the system for this purpose,* ***the overall effectiveness of the system to its user will be the best*** *that is obtainable on the basis of those data."*

Given the relevance of a document as independent to other documents, the ranking relevance score of the document using the probability principle is promised to increase the retrieval effectiveness of the system. Although the PRP assumption unveils toward a new research area in using probability theory to estimate the relevance of documents, it does not specifically determine the way of estimating. Hence several probabilistic models have been proposed with variant methods of estimating the probability of document relevance to a specific query. The general framework of PRP assumption is as follows:

$$Score(q,d) = P(rel/q,d) \approx \sum_{t \in q} wt_{t,d} \tag{2.8}$$

where rel is the relevance probability of a document d given a query q. Robertson and Zaragoza (2009) developed an explicit relevance score for documents derived from a sequence of transformation of the original formula ($P(rel/q,d)$) as ordered-preservation steps. These ordered-preservation steps result in a simple

score formulated as a summation of individual query terms' weights in the document d ($\sum_{t \in q} wt_{t,d}$). In order to evaluate this relevance score, there are two major methods in this probability family, which are binary and best-matching (non-binary) probabilistic models.

2.3.4.1 Binary Independence Model

The Binary Independence Model (BIM) is considered as one of the first probabilistic models that depends on the presence of query terms in a document as a binary relevance of that document to such query. The *independence* concept means that the binary relevance of a query term t to a document d is independent from the other query terms. Given the binary vector of document d as presence or absence of the query terms, the BIM assumption leads us to the following definition score:

$$Score_{BIM}(q,d) = \sum_{t \in q \cap d} \log \frac{P(w_{t,d} = 1/rel_d)(1 - P(w_{t,d} = 1)/\overline{rel_d})}{P(w_{t,d} = 1/\overline{rel_d})(1 - P(w_{t,d} = 1)/rel_d)} \tag{2.9}$$

where rel_d refers to a relevant document d while $\overline{rel_d}$ refers to a non-relevant document d both to the query q. $w_{t,d}$ weight refers to the presence or absence of term t in the document d; which is either 0 or 1.

Equation 2.9 was replaced with the well-known Robertson/Spärck Jones (RSJ) formula (Robertson and Spärck Jones, 1976) in the presence of relevant documents usually from user's feedback as:

$$Score_{RSJ}(q,d) = \sum_{t \in q \cap d} \log \frac{(df_{t,c_{rel}} + 0.5)(N - N_r - df_{t,c} + df_{t,c_{rel}} + 0.5)}{(df_{t,c} - df_{t,c_{rel}} + 0.5)(N_r - df_{t,c_{rel}} + 0.5)} \tag{2.10}$$

where $df_{t,c_{rel}}$ refers to the total number of relevant documents in the corpus that contain the term t and N_r is the total number of relevant documents. Using 0.5 as an additional parameter leads to robust estimation of the results compared to using a simple ratio (Robertson and Spärck Jones, 1976). In a real scenario, the values of N_r and $df_{t,c_{rel}}$ are equal to zero which approximates the formula of $idf_{t,c}$ value in Equation 2.1.

2.3.4.2 Non-binary Retrieval Model

The binary independence model estimates the relevance of documents based on the presence of independent query terms in these documents. This estimation can be improved using user feedback as an indication to determine the relevance of documents. In spite of independence in document relevance values, different documents might have same relevance values if they contain the same query terms regardless of the importance of query terms in a specific document compared to the other documents.

The problem of BIM resides in the binary representation of documents, queries, and relevance, which results in that the model could not determine the documents most likely considered to be relevant to a specific query. Due to this deficiency in BIM, a non-binary term frequency component is proposed to be used in the probabilistic relevance modelling framework (Robertson et al., 1981). Robertson et al. (1981) formulated such framework based on the eliteness concept proposed by Harter (1975) to model the term frequency distribution. The eliteness assumption assigns for each term a set of documents as an elite set which is assumed to be relevant. Hence using the eliteness concept, the frequency of a term in documents could be formulated as a mixture of 2-Poisson distributions for elite and non-elite sets. In particular, given the random variable T of term frequency, the 2-Poisson distribution of the term frequency is as follows:

$$P(T = tf) = \lambda \frac{e^{-\mu_1}(\mu_1)^{tf}}{tf!} + (1 - \lambda) \frac{e^{-\mu_2}(\mu_2)^{tf}}{tf!} \tag{2.11}$$

The first distribution (μ_1) refers to the term frequency in the elite set (somehow relevant to the given term), whereas the second distribution (μ_2) clarifies the term frequency in the rest of the documents in the corpus (non-elite set). The parameter λ refers to the proportion of the documents in the elite set if we assume μ_1 is the population mean of term frequency in the elite set. This can be estimated as a mean of term frequency in this set and in the non-elite set as μ_2.

Although the Harter's 2-Poisson is proposed to estimate the relevance probability of a document for a single term, regardless the weight of a term, it suffers from estimating the parameters in Equation 2.11; λ, (μ_1), and (μ_2) as the relevant documents (elite set) is not available and might face a problem in estimating the relevance for the query of multiple terms.

Robertson and Walker (1994) proposed an effective model that approximates the Harter's 2-Poisson distribution model as a term frequency function of a term t and a document d with the following properties:

$$(a)\ w_{t,d}(0) = 0 \qquad (b)\ w_{t,d}(tf_{t,d}) \propto tf_{t,d} \qquad (c)\ \lim_{tf_{t,d}\to\infty} w_{t,d} = w_{t,d}^{BIM} \tag{2.12}$$

where $w_{t,d}$ refers to weighting scheme for term t in document d where $w_{t,d}^{BIM}$ the term weight using BIM independence model. A single weighting scheme combines these properties can be written with saturation parameter $k > 0$ as:

$$w_{t,d}^{SATU} = \frac{tf_{t,d}}{k + tf_{t,d}} \tag{2.13}$$

As the Harter's 2-Poisson assumes that all documents are in the same length of terms, Robertson et al. (1993) proposed a normalization weight to build a balance between the long documents and the ones that are short as:

$$w_{t,d}^{Norm} = (1 - b) + b(dl/avgdl) \tag{2.14}$$

where dl and $avgdl$ are the length of document d and the average document length of all documents in the corpus (or collection), respectively. The parameter b which is in the interval [0,1] determines the power of normalization as $b = 0$ there is no normalization while $b = 1$ refers to full normalization. In applying this normalisation to the Equation 2.13, a normalised weight scheme is generated as follows:

$$w_{t,d}^{nSATU} = \frac{tf_{t,d}}{k\, w_{t,d}^{Norm} + tf_{t,d}} \tag{2.15}$$

Lastly, Robertson et al. (1994) proposed a ranking function called Okapi-BM25 (or BM25) that combines the Equation 2.15 and Equation 2.10 as follows:

$$Score_{BM25}(q, d) = \sum_{t\in q} w_{t,d}^{nSATU}\, w_{t,d}^{RSJ} \tag{2.16}$$

where $w_{t,d}^{RSJ}$ weight is calculated for the term t and document d which is approximately as Equation 2.1.

Okapi-BM25 is an effective ranking algorithm applied initially on the Okapi system and has been used as a probabilistic retrieval model (Robertson and Zaragoza, 2009). Hence, in this thesis, I will use the BM25 model as retrieval and resource selection method due to its effectiveness and popularity as retrieval model.

2.3.5 Language Modelling

Natural languages comprise of models as a set of rules to generate sentences (or text of terms). Language models use the probability concept to predict the next future term based on the current sequence of terms in a text. In other words, language models use a probability distribution of consecutive terms as a model to predict the next term in the text. In particular, the formal language model of a text (LM_{text}) is a probability function for a sequence of terms $t_1, \cdots, t_n$ given that $text$, which is n-gram (or n sequence of terms) model that predicts a term based on the observed n-1 terms as follows:

$$LM_{text} = P(t_1, \cdots, t_n/text) = \prod_{i=1}^{n} P(t_i/t_1, \cdots, t_{i-1}, text) \tag{2.17}$$

where the projection operator adopts the chain rule while the text parameter could be a query, a document, or a set of documents.

A simple language model predicts the future observation solely on a few previous observed terms, as the entire term observation leads to sparsity for longer terms than the shorter ones. Furthermore, sparsity means that predicting the observed terms yields to predict the observation of their subsequent terms. Due to this problem, the language model should narrow the observed terms to a specific limit $k-1$. Since the n-gram language model of order k is:

$$LM^{k}_{text} \approx \prod_{i=1}^{n} P(t_i/t_{i-(k-1)}, \cdots, t_{i-1}, text) \tag{2.18}$$

This Equation refers to the Markov model (Markov, 1954) of order $k-1$. The most prominent n-gram models are uni-gram ($k = 1$) and bi-gram ($k = 2$) to represent two-term phrases, as follows:

$$LM^{1}_{text} \approx \prod_{i=1}^{n} P(t_i/text) \qquad LM^{2}_{text} \approx \prod_{i=1}^{n} P(t_i/t_{i-1}, text) \tag{2.19}$$

However, language models have been used as a weighting retrieval model in IR systems that use the probability function to estimate how the text (e.g. document d) can generate a given sequence of terms (e.g. query terms). Language models differ from statistical, probabilistic models in Subsection 2.3.4 as the former use a document to estimate its relevance based on relevant queries, whereas the statistical, probabilistic models use a query to estimate the relevance of documents (Zhai, 2008) regarding the distribution of terms as parameter (Ponte and Croft,

1998b); $LM = P(q/d)$ while $PL = P(d/q)$.

Language models are not restricted on constructing a model from a document to generate query terms, but also on constructing a model from a query to generate the document terms. Hence three language retrieval models can be assumed, such as a document language model that generates the probability of query text, a query language model that generates the probability of document text, and a unified language model that combines the two query and document models to predict a joint probability for retrieval. I will discuss the unigram language model under the three language retrieval models as it is sufficient in predicting the text topic and does not depend on the structure of sentences.

In the unigram language model, the order of terms in a text has not an effect on estimating the probability over sequences of terms. The unigram language model is often considered as a bag of words model in which the terms (or words) possess the same probability values despite the different ordering of the terms. Thus, the unigram model could be represented as a multinomial distribution over terms. The multinomial distribution is defined as the probability distribution of the outcomes from multinomial experiment (Tallis, 1962). Moreover, the statistical, multinomial experiment has a set of properties represented as a bag of terms (i.e, document d), which include: (i) the experiment could be repeated over all the terms of a document d with different L_d ordering trials. L_d refers to the number of tokens in the document d. (ii) on each possible ordering of trials, each trial has a discrete number of possible outcomes; which refers to different term frequency of each term in the document d. (iii) the probability of an observed term on each trial is a constant term frequency of that term in the document d. (iv) the trials are independent which means the outcome of one trial does not affect the outcome of other trials. Therefore, to sum over all possible orders of the words, the multinomial probability of a bag of terms has a coefficient in its probability formula as follows:

$$P(d) = \frac{L_d}{tf_{t_1,d}!tf_{t_2,d}!\cdots tf_{t_M,d}!} P(t_1)^{tf_{t_1,d}} P(t_2)^{tf_{t_2,d}} \cdots P(t_M)^{tf_{t_M,d}} \tag{2.20}$$

where L_d denotes the length of a document d (i.e, $\sum_{1 \leq i \leq M} tf_{t_i,d}$) and M is the size of vocabulary (or the number of terms in the document). To reflect that in IR as a model, each document has a unigram language model to estimate the

probability of document to generate a sequence of terms (i.e, query terms). In multinomial Equation 2.20, the coefficient $\frac{L_d}{tf_{t_1,d}!tf_{t_2,d}!\cdots tf_{t_M,d}!}$ is the same for the given query which could be ignored from the calculations.

2.3.5.1 Query Likelihood Model

The first basic language modelling approach in IR is the Query Likelihood Model (QLM) to estimate the probability of generating query terms as a random sample from the documents' language models. In particular, the probability $P(d/q)$ of a document d conditioning on query q represents the QLM models (Ponte and Croft, 1998b). Applying Baye's rule on $P(d/q)$ to get:

$$P(d/q) = \frac{P(q/d)P(d)}{P(q)} \tag{2.21}$$

where $P(q)$ could be ignored as it is equivalent across all documents. Also, the document prior $P(d)$ could be ignored as it is uniformly the same across all documents. The document prior $P(d)$ could be the probability of estimation over other documents' characteristics such as page rank (or authority), length, trustworthiness or others. This simplification leads to a simple probability ranking $P(q/d)$, which is the probability of query q under the language model of document d. Using a multinomial unigram language model, the probability of generating the query terms as sample of text using document d, which is, in turn, the probability of document d, is as follows:

$$Score_{QLM}(q,d) = P_{QLM}(q/LM_d) = \frac{L_q!}{\prod_{t\in q} tf_{t,q}!} \prod_{t\in q} P(t/LM_d)^{tf_{t,q}} \tag{2.22}$$

where $L_q!/\prod_{t\in q} tf_{t,q}!$ is the multinomial coefficient for the query q, which is also a static value for the same query. $P(t/LM_d)$ is the probability of query term t given the document language model LM_d, whereas $tf_{t,q}$ denotes the term frequency of the query term t in the query q.

2.3.5.2 Document Likelihood Model

Document Likelihood Model (DLM) uses the query language model to generate the probability of documents (Lafferty and Zhai, 2001). The assumption is to build a language model from the query side to enhance the query representation

for improving the retrieval quality. The DLM model as QLM is a multinomial unigram language model represented as follows:

$$Score_{DLM}(q,d) = P(d/LM_q) = \frac{L_d!}{\prod_{t \in d} tf_{t,d}!} \prod_{t \in d} P(t/LM_q)^{tf_{t,d}} \tag{2.23}$$

where again $L_d!/\prod_{t \in d} tf_{t,d}!$ is the multinomial coefficient for the document d, which is also a static value for the same document over all the queries. $P(t/LM_q)$ is the probability of term t in document d given the query language model LM_q, whereas $tf_{t,d}$ denotes the term frequency of the query term t in the document d. Applying the DLM approach on IR systems results in an inefficient estimation, as the language model built over query is not enough as in the document language model (Lafferty and Zhai, 2001). Due to this problem, the query language model could be enhanced from other models such as relevance feedback from the user or pseudo-relevance feedback from the top retrieved document terms to expand the query terms and then generate an accurate query language model (Lavrenko and Croft, 2001).

2.3.5.3 Unified Likelihood model

A unified query and document language model formulation is an effective approach that combines the two language models into a ranked-based function in IR systems. In particular, Lafferty and Zhai (2001) developed a risk minimization approach for retrieving the document d as relevant to the query q in the language modelling framework. The risk of retrieving documents, that their language models do not match the query language model, is quantified as a Kullback- Leibler (KL) divergence between the document and query language models, as follows:

$$Score_{unified}(q,d) = KL(LM_d||LM_q) = \sum_{t \in V} P(t/LM_q) \log \frac{P(t/LM_q)}{P(t/LM_d)} \tag{2.24}$$

Asymmetric KL divergence as in information theory measures how the LM_d represents the probability distribution LM_q (Manning and Schütze, 1999). In comparison with the previous likelihood models, unified likelihood model shows more effective results (Lafferty and Zhai, 2001) and also is represented as the current state-of-the-art method in language modelling of IR systems (Zhai, 2008).

2.3.5.4 Term-based Language Model Estimation

The main component in the language model approaches is how to estimate the probability of generating query terms from a specific language model (i.e, $P(t/LM_{text})$) as a language model process. The simplest and commonly used method is the Maximum Likelihood Estimation (MLE; (Fisher, 1922)); defined as:

$$LM_{text} = P_{MLE} = \frac{tf_{t,text}}{L_{text}} \tag{2.25}$$

where $tf_{t,text}$ is the term frequency of term t in a sample $text$ (i.e, query or document) while L_{text} is the number of tokens in the $text$. As an example, using the MLE method over all query terms in unigram language modelling framework gives the following formula:

$$P(q/LM_d) = \prod_{t \in q} P_{MLE}(t/LM_d) = \prod_{t \in q} \frac{tf_{t,d}}{L_d} \tag{2.26}$$

However, the sparseness of appearing the terms in documents leads to a major problem in the language modelling approaches. The language model assigns zero probability to the query if at least one of its terms does not appear in the document d, although the rest of the terms are contained in the document (Manning et al., 2008a). In order to tackle such deficits, an effective smoothing approach is needed to balance the non-zero probabilities as weighting components by giving a bit probability to unseen terms to avoid the zero probability of terms (Zhai and Lafferty, 2004).

Several smoothing methods have been proposed. One of the simplest approaches is to combine the query and document language models in a linear smoothed interpolation function (or Jelinek-Mercer smoothing) with the general corpus language model (LM_c) (Hiemstra, 2001), which is estimated as follows:

$$P_\lambda(t/text) = \lambda P_{MLE}(t/text) + (1 - \lambda) P_{MLE}(t/C) \tag{2.27}$$

where $0 \leq \lambda \leq 1$ is the interpolation parameter and C refers to the entire corpus language text.The entire corpus language model is promised to solve the zero-probability problem as the term has a high chance to appear in this corpus. Another alternative and special case of the linear smoothed interpolation approach is to use the Bayesian smoothing with Dirichlet prior that is updated

from the Bayesian process with parameter μ (Mackay and Peto, 1994), which is defined as:

$$P_{\mu}(t/text) = \frac{tf_{t,text} + \mu P_{MLE}(t/C)}{L_{text} + \mu} \tag{2.28}$$

In this Equation, $\lambda = \mu/(L_{text} + \mu)$ is the state-of-the-art performance of the Dirichlet smoothing method (Zhai, 2008).

2.3.6 Divergence from Randomness

Divergence From Randomness (DFR) is a probabilistic, query-dependent ranking approach that is built on an assumption that the informativeness of a document depends on the deviation of its terms' frequencies distribution from random distribution (Amati, 2003). Hence DFR models, similar to the best matching models in the Subsection 2.3.4.2, is derived from the Harter's 2-Poisson assumption, in which the informativeness of term in a corpus can be analysed by its distribution on a set of documents (Harter, 1975). Particularly, in DFR models, the frequency of informative terms tends to be more in a few elite set while the frequency of non-informative terms is distributed randomly over the document corpus. This phenomenon can be modelled by a Poisson distribution as an average (or mean) frequency of the terms in the corpus. However, DFR models differ from the best matching and language models in exploiting the statistical distribution of the documents' term frequencies and the implicit account of relevance. The hypothesis behind the DFR models -assuming that the elite set of a term is the set of documents that contain the term (Amati and Rijsbergen, 2002)- is that "the informative content of a term can be measured by examining how much the term frequency distribution departs from a benchmark distribution, that is, the distribution described by a random process " (Amati, 2003). A quantitative DFR model is formulated as follows:

$$Score_{DFR}(q,d) = \sum_{t \in q} w_{t,q} * w_{t,d} = \sum_{t \in q} \frac{tf_{t,q}}{Max_{t_i \in q} tf_{t_i,q}} * (-\log_2 p1(t/C)(1 - p2(t/d))) \tag{2.29}$$

where $-\log_2 p1(t/C)$ and $(1 - p2(t/d))$ represent the informativeness of the term t in a corpus C and document d that contain t, respectively. The probability $p1(t/C)$ is a basic randomness model of the distribution of term t in the corpus C such as divergence approximation of the binomial, approximation of the binomial,

bose-einstein distribution, geometric approximation of the bose-einstein, inverse document frequency model, inverse term-frequency model, and inverse expected document frequency model (Amati, 2003). The probability $p2(t/d)$ defines the information gain of observing the term t in the document d, which can be computed using two models; Laplace L model ($\frac{1}{tf_{t,d}+1}$) and the ratio of two bernoulli's processes B model ($\frac{F}{df_{t,c}.(tf_{t,d}+1)}$). However, a term frequency normalization is needed as a third component in DFR models; $tf_n = tf_{t,d}.\log_2(1+c.\frac{avgdl}{dl})$, which is called normalisation 2 and c is a free parameter. This normalisation is due to the amount of information in a document that is proportion to its length. Since a variation of functions of the three components can define different effective DFR models (Amati, 2003).

In this thesis, I use the DFR models that are implemented in Terrier open source information system[1] and developed by Amati and van Rijsbergen framework (Amati and Rijsbergen, 2002). I briefly mention some of these models as follows:

1. BB2 model uses a Bernoulli-Einstein with Bernoulli after-effect and normalisation 2. The $w_{t,d}$ weight of the BB2 model is calculated as Equation 2.30.

$$\begin{aligned} w_{t,d} &= \log_2 \frac{(N+F-tf_{t,d}-2)!F!(N-1)}{(F-tf_{t,d})!(N+F-1)!} \\ &= \frac{F+1}{df_{t,c}\cdot(tfn+1)}\big(-\log_2(N-1)-\log_2(e)+f(N+F-1,N+F-tfn-2)-f(F,F-tfn)\big) \end{aligned} \tag{2.30}$$

2. InL2 uses an Inverse Document Frequency model with Laplace after-effect and normalisation 2. The $w_{t,d}$ weight of the InL2 model is calculated as Equation 2.31.

$$w_{t,d} = \frac{1}{tfn+1}\big(tfn\cdot\log_2\frac{N+1}{df_{t,c}+0.5}\big) \tag{2.31}$$

3. In_expB2 uses an inverse expected document frequency model with Bernoulli after-effect and normalisation 2. The $w_{t,d}$ weight of the In_expB2 model is calculated as Equation 2.32.

$$w_{t,d} = \frac{F+1}{df_{t,d}\cdot(tfn+1)}\big(tfn\cdot\log_2\frac{N+1}{df_{t,c}+0.5}\big) \tag{2.32}$$

[1] http://terrier.org/ (October, 2016)

4. In_expC2 uses an inverse expected document frequency model with Bernoulli after-effect and normalisation 2. The $w_{t,d}$ weight of the In_expC2 model is calculated as Equation 2.33.

$$w_{t,d} = \frac{F+1}{df_{t,c} \cdot (tfn_e + 1)} \left(tfn_e \cdot \log_2 \frac{N+1}{df_{t,c} + 0.5} \right) \tag{2.33}$$

where $tf_{t,d}$ is the within-document frequency of t in d. $avgdl$ is the average document length in the collection. dl is the document length of d, which is the number of tokens in d. N is the number of documents in the collection. F is the term frequency of t in the whole collection. $df_{t,c}$ is the document frequency of t in corpus c. tf_n is the normalised term frequency. tfn_e is also the normalised term frequency. It is given by a modified version of the normalisation 2:

$$tfn_e = tf \cdot \log_e(1 + c \cdot \frac{avgdl}{dl}) \tag{2.34}$$

n_e is given by $N(1-(1-nt/N)F)$. The relation f is given by the Stirling formula as follows (Feller, 1968):

$$f(n,m) = (m+0.5)log_2\frac{n}{m} + (n-m)log_2 n \tag{2.35}$$

2.3.7 Divergence from Independence

A closely related model to DFR is Divergence From Independence (DFI) model (Dinçer, 2012), which is a non-parametric weighting model that quantifies the importance of documents to specific terms by the amount of divergence of terms' frequencies in documents *independently* from the expected frequencies in the corpus. Hence the difference between DFI and DFR models is that DFI model estimates the importance of a given document to a specific query terms by calculating the divergence of the query terms' frequencies from the frequencies suggested by an independence model while DFR by the randomness model. In particular, the DFI score of a term t_i in a document d_j is estimated as the difference of the frequency of the term in that document (tf_{t_i,d_j}) from the expected frequency e_{ij} where e_{ij} is as follows:

$$e_{ij} = TF_{t_i,c} \cdot \frac{dl_{d_j}}{\sum_{k=1}^{N} dl_{d_k}} \tag{2.36}$$

where $TF_{t_i,c}$ is the total frequency of term t_i in the corpus c and dl_{d_j} refers to document length of d_j. Hence the DFI score of term t_i to document d_j is as

follows:

$$DFI_{t_i,d_j} = \log_2\left(\frac{tf_{t_i,d_j} - e_{ij}}{\sqrt{e_{ij}}}\right) \tag{2.37}$$

In practise, if $DFI_{ij} \leq 0$ the value will be zero.

2.3.8 Learning to Rank Approaches

Information retrieval models require manually tuning parameters as a validation process to effectively estimate the relevance of documents to the user's information need (HE and Ounis, 2003; Taylor et al., 2006). Due to the necessity for automatic ranking systems to alleviate the problem, Learning to Rank (LtR) approaches have emerged as supervised machine learning algorithms that automatically tune these parameters through combining different relevance estimators in a best fitting model for document ranking task (Cao et al., 2007). Figure 2.2 describes the learning and prediction process in LtR approaches.

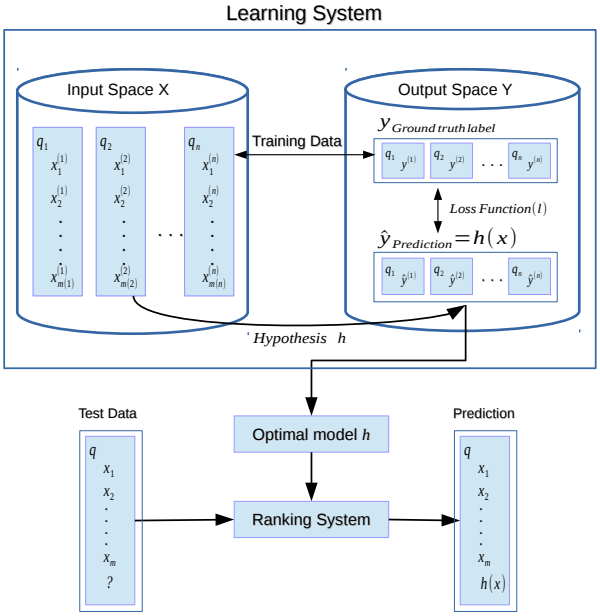


Figure 2.2: Learning To Rank Framework (Liu, 2011)

Liu (2011) narrows the concept of LtR approach as a ranking method of feature-based and discriminative training. The feature-based concept means that the input space has a set of feature vectors reflect the relevance of documents to a

specific query. Discriminative training concept means that the learning process, which contains four components; input space (X), output space (Y), hypothesis space (H)), and loss function (l), occurs at each query and then the optimisation is run over all the queries. In particular, as seen in Figure 2.2, the input space (X) consists of m number of document features as vector $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \cdots, x_1^{(m(i))}\}$ associated with a query q_i and y^i as relevance judgement grade in output space (Y), where $i = 1, ..., n$ is used to form a training data in the learning system. The features can be different relevance estimators of document vary from query dependent features such as retrieval model scores to query independent features such as page rank score while the graded relevance label can be real, ordinal, or nominal types of data. The hypothesis space H uses a scoring function to build a set of hypotheses through using a loss function (l) and finds the best hypothesis h for the testing phase. The loss function computes the inconsistency measure between the prediction $\hat{y}$ and the ground truth label y mapped from input space X and output space Y.

LtR algorithms can be categorised into three groups, which are pointwise approaches (Friedman, 2000; Breiman, 2001), pairwise approaches (Burges et al., 2007; Wu et al., 2008), and listwise approaches (Xu and Li, 2007b; Metzler and Bruce Croft, 2007). The pointwise approaches predict, using a scoring function, the relevance grade of the input feature vectors of each single document and then sort these scores of all documents in a final ranked list. The pointwise approaches are essentially traditional machine learning algorithms, which could be classified as classification (not quantitative values) (Li et al., 2008), regression (real values) (Cossock and Zhang, 2006), and ordinal regression (ordinal) (Sorokina et al., 2007). The pointwise approaches do not take into consideration the interdependency between documents at training phase where each document is independent from other documents in ranking (Liu, 2011). In the pairwise approach, the input space is a pair of documents and the output space is the grade relevance that reflects the preference of a document to the other one. The loss function in pairwise approaches finds the hypothesis that minimises the number of miss-re-ordered (or miss-classified) document pairs in the ranking. The limitation of pairwise approaches reside in difficulty deriving the position of documents in ranked list using pair of documents in ranking. Hence the pointwise and pairwise

approaches ignore the fact that the same documents or pair of documents are associated to the same query (Cao et al., 2007). In contrast, the listwise approaches treat a permutation of a set of documents as a basic unit (a group of documents) along with a specific query in the training phase and builds loss functions in this form. Then the listwise model sorts the relevance of documents for a query as an ordered rank list. In particular, the listwise approaches predict with a scoring function the graded relevance of feature vectors related to a list of documents of the same query. The listwise approaches aim sometimes to optimise the value of a particular information retrieval evaluation metric, averaged over all queries in the training data. In this case, however, the loss function considers the positions of the documents in the ranked list of all the documents associated with the same query (Liu, 2011). I use the following LtR approaches in this thesis due to the importance of these algorithms in the literature research (Liu, 2009).

Random Forest (Breiman, 2001) is a pointwise boosting ensemble approach that builds a set of decision trees as de-correlated models with low correlation on their predictions. In particular, at each iteration of boosting, a set of random instances and features from the training set is drawn with replacement and a decision tree is built on that random sample. The prediction on new data occurs through testing the majority vote (classification) or average value (regression) on all generated trees.

MART (Friedman, 2000) is a pointwise multivariant, gradient boosted (deepest descent) regression tree algorithm that builds a ranking model as a weighted linear combination of an ensemble of weak regression learner trees (i.e, boosting technique) to fit the training output to their relevance scores. The MART algorithm builds each weak learner in a stepwise fashion to be fitted to the gradient (or pseudo-residuals) of the previous model at each leaf node of the regression tree to estimate the weighted parameters.

RankNet (Burges et al., 2005) is a pairwise artificial neural network model that minimises the misclassified query-object pairs by estimating the difference between a pair of network outputs as a probabilistic cross entropy cost function. Thus, the neural network model ranks the future queries based on the evolved model trained on the cross-entropy cost function and adapted modified back-propagation algorithm.

LambdaRank (Burges et al., 2007) is a listwise approach that approximates the gradient of IR measurement cost by modelling the gradient of each document pair in the dataset with lambda function, called $\lambda-$gradients. In particular, the optimisation reduction in pairwise cross entropy loss function used with RankNet is not suitable to improve the ranking performance of IR evaluation metrics such as nDCG, MAP, MRR, etc[1]. As the IR measurement metrics are discrete and not differentiable, the gradient of the metric cannot be optimised directly. LambdaRank estimates the gradients by scaling the RankNet cross-entropy function with the amount of gain in nDCG metric (or other IR metrics) by swapping the two documents in the list.

LambdaMART (Burges, 2010) is a listwise learning to rank algorithm that combines the gradient boosting optimisation in MART algorithm and the listwise LambdaRank model to optimise directly measure-specific cost function such as nDCG metric. In particular, LambdaMART uses the lambda gradient idea in the LambdaRank model to be used as a gradient loss function in MART algorithm as gradient boosting regression trees model.

Coordinate Ascent (CA) (Metzler and Bruce Croft, 2007) is a listwise unconstrained linear optimisation model that iteratively optimises a multivariant objective function through a series of one-dimensional line search. It cycles repeatedly through each parameter and optimises over it while holding the others fixed until no improvement is observed. Specifically, the algorithm optimises through minimisation of a measure-specific loss function, which is called Mean Average Precision (MAP) as an information retrieval evaluation metric. Although the algorithm suffers from getting stuck in local minima when searching for the global minima of the MAP, it has been proven to be highly effective for a small number of parameters and has good empirically verified generalisation properties (Bendersky et al., 2010).

AdaRank (Xu and Li, 2007b) is a listwise boosting algorithm that optimises a specific IR performance measure. In particular, AdaRank builds a final learner iteratively through combining a weak ranker at a time. Each weak learner has a weight and ranking model where the weight is updated by increasing the weights of resources (or documents) not being ranked properly using created model and

[1]I will discuss a set of evaluation methods in the next Section 2.4

focusing merely on weak rankers that mistakenly ranked the objects.

RankBoost (Freund et al., 2003) is a pairwise boosting approach that builds a ranking model by combining weak rankers on a set of iteration one ranker per iteration. The algorithm attempts to minimise the number of mis-classification input pairs of query-object. Hence RankBoost approach plugs the exponential loss of document pairs into a framework of AdaBoost technique (Freund and Schapire, 1997), which works as an AdaRank algorithm in creating the ranker but in a pairwise manner.

2.4 Information Retrieval Evaluation

The effectiveness and the efficiency of an information system can be assessed and improved using an essential task called retrieval evaluation. Subsection 2.4.1 clarifies the evaluation methodology, while Subsection 2.4.2 explains the evaluation metrics.

2.4.1 Evaluation Methodologies

IR researchers have focused on building an effective and efficient IR system that satisfies a user's information need(s). In order to achieve such a goal, an evaluation task is used to assess the ability of the system in providing relevant information within sufficient response time.

IR evaluation essentially depends on two experimental methodologies: System-oriented evaluation and User-centred evaluation (Croft et al., 2009; Baeza-Yates and Ribeiro-Neto, 2011). System-oriented evaluation is a well-designed methodology providing an objective comparison between retrieval systems (Cleverdon et al., 1966; Borlund, 2003). The birth of system-oriented evaluation dated back to the earliest large-scale evaluation experiment in 1960s (Cleverdon et al., 1966), which is called the Cranfield evaluation paradigm. In this paradigm, researchers create a controlled test collection or evaluation corpus (Voorhees and Harman, 2005) as a benchmark of documents, queries (or topics), and relevance assessment judgements (or qrel) in addition to two measurement ratios; which are precision and recall as I will clarify in the next Subsection 2.4.2. Using this approach, two systems are fairly compared with respect to effectiveness on the same ground

truth of evaluation corpus. This comparison occurs through issuing a query to both systems and then the relevance of retrieved documents on two result lists are separately evaluated using the qrel (**q**uery **rel**evance) file. The relevance of documents in the qrel file is assessed by a set of experts (or assessors) through giving them the queries and getting their judgements on these documents, which are usually binary assessment judgement such as relevant or not relevant. Given the qrel file, the measurement metrics (such as Precision and Recall) are used to assess the effectiveness of both systems based on the relevant retrieved documents. As an example, two evaluation campaigns adopt the Cranfield paradigm as a standard approach for IR evaluation, which are **T**ext **RE**trieval **C**onference (**TREC**) (Voorhees and Harman, 2005) and **C**ross **L**anguage **E**valuation **F**orum (**CLEF**) (Gey et al., 2005; Ferro, 2014). Recent has witnessed a large set of modern test collections such as Common Crawl Corpus[1] (Approx. 5 billion English documents) and ClueWeb12[2] (Approx. 733 million English documents). Due to the time-consuming and effort cost in judging such collections, the relevance assessment usually is estimated over a subset of the collection for each query. Different IR systems are used to gather top k diverse results for each topic and the results are aggregated into a merged list of results (or pool of results) for judging. The relevance assessors judge the relevance of each document in the pool which can then be used to compute system effectiveness where the rest of documents often are deemed to be non-relevant even if they are relevant in a standard approach called pooling (van Rijsbergen, 1979; Buckley et al., 2007; Webber and Park, 2009). A recent approach for judging a set of documents is called crowdsourcing (Carvalho et al., 2011). Crowdsourcing is the process of sourcing the documents to be evaluated to a large crowd of online users each conducts a small evaluation task in the form of an open call (Clough et al., 2013). As an example Amazon Mechanical Turk (AMT)[3] is a crowdsourcing platform of around 822,969 workers from different countries that perform human intelligence tasks. One of the recent evaluation methods is A/B testing method that is derived from the controlled experiments framework that involve real users to evlaute a new idea

[1] https://aws.amazon.com/public-datasets/common-crawl/ (October, 2016)
[2] http://www.lemurproject.org/clueweb12.php/ (October, 2016)
[3] https://www.mturk.com/mturk/welcome (October,2016)

of an online system (Kohavi et al., 2013). The users are distributed randomly into two groups of the original (A) and the updated (B) versions of the system and then the controlled experiment monitors the behaviour of the users on both groups using specific metrics. The statistically significant between two groups is conducted on the observed differences results to evaluate the random noise for a decision on changing the current system. On the same phenomenon, the IR A/B evaluation, the relevance of documents depend on user actions (e.g, clicks) per query level of two ranking results (Bailey et al., 2007).

System-oriented evaluation is not sufficient for evaluating interactive IR systems (Robertson and Hancock-Beaulieu, 1992). Since the user-centred or task-oriented evaluation is used to involve the users in the evaluation process through interactive IR systems to include the cognitive behavioural features of their decision on the relevance assessments over time. This is due to the fact that information needs change with new information in search results that is varying with different users (Ingwersen and Järvelin, 2005; Belkin, 1980). User-centred evaluation is used for estimating the effectiveness of interactive IR systems (Borlund, 2003). In this thesis, I will focus on system-oriented evaluation in P2P-IR systems due to the limited time.

2.4.2 Evaluation Metrics

The evaluation methodologies present a benchmark of the test collection to assess the effectiveness of IR systems. The effectiveness measure can be estimated using several metrics to quantify the ability of the retrieval system in providing more relevant documents at cut-off value k of retrieved documents. The evaluation metrics categorised into two classes, which are set-based metrics and position-based metrics. The set-based metrics concern the number of relevant documents whilst position-based approach depends on the position of relevant documents in the evaluation. In this section, I will discuss the evaluation metrics that is used in this thesis.

2.4.2.1 Set-based Evaluation Metrics

The set-based metrics are considered some of the most basic metrics with situations that do not take into account the relevant documents' positions in evaluation

using the ranked results list. There are three prominent set-based metrics, which are Precision, Recall, and F-measure. Precision refers to the number of relevant documents to the retrieved documents in the result list. Recall denotes the number of relevant documents in the result list to the whole relevant documents of a specific query. Given the ranked list R_q of a query q of retrieved documents and Rel_q of the whole relevant documents of the query q, the metrics Precision and Recall are defined as follows:

$$Precision@k = \frac{|Rel_q \cap Ret_q^k|}{|Ret_q^k|} \qquad Recall@k = \frac{|Rel_q \cap Ret_q^k|}{|Rel_q|} \tag{2.38}$$

where Ret_q^k is the top k documents retrieved for the query q. The two metrics can be estimated at any cut-off values of k, but the basic measures calculated on the whole retrieved result list. Kent et al. (1954) is the first who introduced the Precision and Recall concepts and then analysed by the IR community (Cleverdon, 1972; Raghavan et al., 1989; Salton, 1971). An important property is the inverse relationship between the two metrics as the Precision increases whilst the Recall decreases on the increasing number of retrieved documents.

F-measure combines the Precision and Recall values in one metric in order to balance the two metrics and tune them for effectiveness evaluation. van Rijsbergen (1979) proposed the F-measure as an evaluation metric that combines the Precision and Recall metrics through tuning a parameter called $\beta \in [0, \infty]$ as follows:

$$F_\beta = \frac{(\beta^2 + 1).Precision.Recall}{\beta^2.Precision + Recall} \tag{2.39}$$

A harmonic mean of Precision and Recall is through using the default value of β which is 1 as $F_{\beta=1}$.

2.4.2.2 Position-based Evaluation Metrics

The set-based metrics count the number of relevant documents at cut-off value leaving into consideration the importance of relevant documents at specific positions (Robertson, 2008). This limitation discriminates the effectiveness of the retrieved results of two retrieval systems that have the same number of relevant documents at cut-off point with different ranks. Hence a set of position-based metrics solve the problem where in this thesis I used the following methods:

Average Precision (AP; Harman (1993)). The AP metric is defined as the

average number of Precision values at each position of the relevant retrieved documents.

$$AP@k = \frac{\sum_{i=1}^{k} Precision@i \times Rel_i}{|Rel_q|} \tag{2.40}$$

where $Precision@i$ is the Precision at the i-th position, while Rel_i is an indicator equal one if the document at that position is relevant, zero otherwise. This AP is then averaged over a set of queries to generate the Mean Average Precision (MAP) as:

$$MAP@k = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{1}{|Rel_q|} \sum_{i=1}^{k} Precision@i \times Rel_i \tag{2.41}$$

where q_i refers to a query from a set of queries Q. MAP reflects the system effectiveness over all queries in a benchmark of the test collection.

Discounted Cumulative Gain (DCG; Järvelin and Kekäläinen (2002)). The relevant documents differ in the level of relevance to a specific query (Teevan et al., 2007), since the binary assumption of relevance is limited. Järvelin and Kekäläinen (2002) suggested a graded scale to evaluate the relevance of documents from less relevant to more relevant. Therefore, a log-based discount factor is proposed to the higher relevant documents in evaluation over the lower relevant ones. The evolved metric called DCG which is defined as:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{Relevance_i} - 1}{\log_2(i+1)} \tag{2.42}$$

where $Relevance_i$ is a non-binary relevance grade related to a document at the i-th position. However, different relevance grades can be used as in web scenario that range in five values (Burges et al., 2005) and different logarithmic bases can be used for large and small discounts (Järvelin and Kekäläinen, 2002).

$nDCG$ is a popularly derived effectiveness measure estimated by normalising the $DCG@k$ with the maximum possible (ideal) $IDCG$ value for a given set of queries on the ideal ranked list as follows:

$$nDCG@k = \frac{DCG@k}{IDCG@k} \tag{2.43}$$

Expected Reciprocal Rank (ERR). ERR is a novel metric that depends on the cascading user model (Craswell et al., 2008), which assumes the expected usefulness of documents based on a cascading model of user browsing behaviour

(Chapelle et al., 2009). In the cascading model, the user scans ranked documents from top to bottom and has a specific probability for each document of being satisfied and useful. However, the usefulness of a currently scanned document depends on unsatisfied previously viewed documents. In other words, ERR is just the expectation of the reciprocal of the position of a result at which a user stops. Intuitively, according to this cascading browsing model, the high probability that the user stops scanning more documents once he/she found the relevant ones. To formalise this, let P_r denote the relevant probability of a document at rank r to the query, and let $\prod_{i=1}^{r-1}(1-P_i)$ denote the probability that the user is not satisfied with documents from ranks 1 to $r-1$. Then, ERR is defined based on the expected probability that the user is finally satisfied at rank r as follows:

$$ERR@r = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1}(1-P_i)P_r \tag{2.44}$$

In practice, P_i is defined as a function of the relevance grade g_i of the i^{th} document, i.e., $P_i = (2^{g_i}-1)/2^{g_{max}-1}$, where g_{max} is the maximum considered grade.

2.5 Relevance Information Feedback

The user on the web issues a query using search engines to find the relevant information that meet his/her information need. The given query might not be adequately informative to retrieve a high quality results, as it is short in length and do not reflect the user information need (Phan et al., 2007). Therefore search engines use relevance feedback in their query processing component as an information retrieval feature in order to enhance the issued query and then to improve the quality of retrieved results.

The main idea behind relevance feedback is to exploit the initial result of a query and explores whether this retrieved information is relevant for enhancing the query to match the user information need accurately (Baeza-Yates and Ribeiro-Neto, 2011). The relevance feedback reformulates the given query in two forms; like adjusting the weights of the query terms or adding more words to the query terms (van Rijsbergen, 1979; Croft et al., 2009). There are three different types of relevance feedback for improving the quality of retrieved results, which

are explicit feedback, implicit feedback, and blind (or pseudo) feedback (Manning et al., 2008b). In the explicit relevance feedback, the assessors (or users) provide the relevance (binary or graded relevance feedback) to the retrieved documents of the given query where they know that the provided feedback is necessary for the benefit of IR systems and also is deemed as a relevance judgement (Ruthven and Lalmas, 2003). The implicit relevance feedback is derived from user behaviour during searching sessions where the user satisfies his/her information need without assessing the relevance for the benefit of the IR system. The user behaviour includes (i) clicking through data (or selecting documents for viewing) (Joachims, 2002), (ii) dwell time that refers to the duration time the user spent to view a document (Kim et al., 2014), and (iii) document browsing or scrolling actions (Kelly and Belkin, 2001). Blind (or pseudo) relevance feedback provides an approach that analyses the top k (k between 10-50 in most experiments) retrieved documents assuming their relevancy to the initial query (van Rijsbergen, 1979; Shipeng Yu, 2002; Ruthven and Lalmas, 2003). Using a specific weighting scheme, the blind feedback method selects candidate terms from the top documents to expand the given query. The expanded query then is used again to retrieve the most relevant documents that might be missed in the initial search.

The IR systems, in order to use the implicit and explicit relevance feedback in simulation, have to use a large number of assessors (or users) to provide explicit feedback or monitor the behaviour of users to get the relevance information. In spite of difficulty in obtaining the information feedback in such scenarios, implicit and explicit feedback can be simulated (Maxwell and Azzopardi, 2016). Recent developments in the area of Interactive Information Retrieval (IIR) have seen the introduction of complex user simulations. These simulations have reached a stage such that the underlying frameworks can be modified and adapted such that simulations can now provide levels of explicit (or implicit, if models permit) feedback as required.

The development of *autonomous search agents* (Maxwell and Azzopardi, 2016; Maxwell et al., 2015) is one such exciting development in the field of IIR simulation. These agents are highly configurable, credible abstractions of real-world users utilising a search engine. They can issue multiple queries and judge snippets and documents autonomously - that is, without recourse to any prior relevancy

information. The agents have demonstrated a marked improvement in terms of mirroring the behavioural characteristics of real-world searchers under a specific search task when compared the current state of the art (e.g. Baskaya et al. (2013); Maxwell et al. (2015)).

These agents could in theory be used as part of a test harness for a distributed IR system, especially for peer-to-peer networking. Agents could be utilised as simulated users, and thus drive the system, issuing queries, examining documents, and judging documents for relevance. The main advantage of using agents in this way is that they can be highly controlled. As an example, a particular configuration can be employed in a 'what-if' scenario, allowing one to closely observe the performance of the distributed system, knowing that the inputs to the system are being issued in a controlled manner.

2.6 Advantages and Disadvantages of Search Engines

The search engine infrastructure might be built under Internet or enterprise private network. The search engine applications reside in a centralised powerful, robust, and enough space storage machines called servers to allow other machines called clients find their information need. Although search engines have advantages of simplicity in document management and high efficiency in comprehensive search and retrieving information of surface web, they are susceptible to various deficiencies; for instance, due to high cost of storage space and computing power, search engines companies are monopoly on the information and have full control over them (Kulathuramaiyer and Balke, 2006; Mowshowitz and Kawaguchi, 2002), search engines leave users prone unethically to privacy risk by pursuing their behaviours (Tene, 2009), search engines have to dynamically keep tracking the updated information on Internet (Lewandowski et al., 2006), and finally crawlers in search engines might be unable to locate web pages called the hidden web (or deep web) that are invisible to be indexed and accessed by at least one person (Bergman, 2001; He et al., 2013; Rudesill et al., 2015).

2.7 Chapter Summary

This chapter discusses a survey and background information about the web search engines including their components, ranking models, evaluation methodologies, and relevance information feedback that will be used throughout the thesis. In addition, I explained the search engines' advantages and disadvantages. In each section, I clarified the work that is related to the thesis contribution, which is enhancing the query routing in semi-structured P2P-IR networks. In particular, I explained crawling, indexing, and query processing components of the web search engines to highlight the problem of crawling process in extracting the hidden web for indexing and manifest the indexing and query processing components that are used in P2P-IR networks. The information retrieval models are discussed using different families including Learning to Rank (LtR) algorithms. These retrieval models will be used to retrieve relevant documents and will be adapted as resource selection methods in the target semi-structured P2P-IR networks. In particular, the retrieval models are used as retrieval and ranking approaches in distributed resources (or peers) and resource selection methods at super-peer level. The evaluation methodologies are also clarified to evaluate the retrieval effectiveness of IR systems. The evaluation methodologies use evaluation metrics to estimate the retrieval quality according to the number of relevant documents at the top of the result list. I presented a set of metrics that are used in this thesis for retrieval perspective and as evaluation metrics for loss functions in LtR algorithms throughout the thesis. The implicit feedback information simulated from the user interaction will be utilised to enhance the query routing in semi-structured P2P-IR network as discussed in Chapter 6, hence I discussed the relevance information feedback and explained the importance of simulating the user information feedback for improving the retrieval results. Finally, I outlined the advantages and disadvantages of search engines that motivate to propose a promising technology to tackle their problems and improve the quality of retrieval results.

Chapter 3

Distributed Information Retrieval

> "The strong person is not the good wrestler. Rather, the strong person is the one who controls himself when he is angry."
>
> — Prophet Muhammad, (570-632 A.D)

3.1 Introduction

Searching the web for relevant documents is considered one of the most popular tasks on the Internet. Search engines provide an effective retrieval performance for efficiently storing and retrieving documents. Although search engines have advantages in efficiently retrieving relevant documents, they have ethical and technical drawbacks as discussed in Chapter 2. Federated Search (or Distributed Information Retrieval) systems have emerged as a promising paradigm to alleviate the search engine problems. However, federated search systems provide a uniform interface across a plurality of searchable resources by way of a broker. The broker submits a query in parallel to these resources (or text collections) that have a high probability of relevant documents. The retrieved result lists of selected collections are then merged into a final result list for users to meet their information need as shown in Figure 3.1.

There are three forms of federated search systems (Shokouhi and Si, 2011): meta-search, vertical (or aggregated) search, and Peer to Peer (P2P) network search. In meta-search[1][2][3], the broker sends a given query in parallel to multiple

[1] http://www.dogpile.co (October, 2016)
[2] http://monstercrawler.com/ (October, 2016)
[3] http://www.allinonenews.com/ (October, 2016)

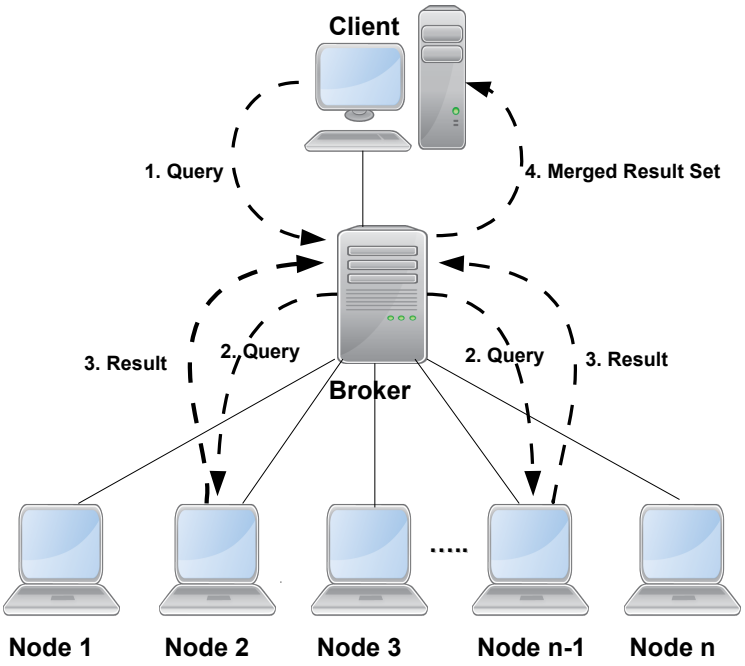


Figure 3.1: Federated search Diagram

search engines and merges the retrieved result lists into a final ranked list (Meng et al., 2002). In vertical search[1], the broker sends a query to a set of search verticals (e.g, images, news, blogs, books, videos, and maps) of different topics and incorporates the retrieved multimedia answers along with the merged default text search results into a final ranked list (Hawking, 2004; Bailey et al., 2007; Kopliku et al., 2014). P2P network search is considered as a complicated form of the federated search systems, in which a query is sent to multiple resources (or peers) at different points of nodes and the retrieved result lists are merged along the path from the responded resources to the query sender (or customer) into a final list (Tigelaar et al., 2012; Klampanos and Jose, 2012). These networks might contain a broker or multiple brokers to manage the process of searching due to facilitate routing query between the self-organising peers and merging the results on behalf of their peers in the system. Because it is the main focus of this thesis, I survey the P2P network search in more detail in Section 3.4.1.

[1] http://www.naver.com (October, 2016)

Due to the distributed nature of the resources in federated search, there are three challenges that have to be tackled, which include resource representation, resource selection, and result merging (Callan, 2000). Resource representation concerns the acquisition of sufficient and representative information for ranking and selecting decisions, whereas more accurate information leads to more accurate, effective, and efficient selection decision. The resource selection process utilises the acquired representations as an indication of relevance for ranking resources, whereas the resources with high relevance to a given query, and most likely contain relevant documents, are selected to answer the query. Result merging combines the retrieved result lists of the target resources into a final result list. However, various approaches have been exhibited and differ in presenting the resources, locating them, and merging the results in two environments of federated search. The two environments depend on whether the resources are cooperative or uncooperative (Callan, 2000; Crestani and Markov, 2013). In cooperative environments, the brokers acquire meta-data statistical information from their resources to make a decision regarding routing a given query to those most likely to contain relevant documents (Callan et al., 1995; Gravano et al., 1999; Yuwono and Lee, 1997; Xu and Croft, 1999; Melucci and Poggiani, 2007; Aly et al., 2013). On the other hand, in uncooperative environments, the brokers gather offline the required information that is a sample of documents usually by randomly issuing sample queries and collecting top matching documents from multiple resources in a process called query sampling method (Callan and Connell, 2001). The broker then creates a centralised sample index (CSI) from the sampled documents as a representative information of these resources. By running the given query on CSI, the broker selects the relevant ones that have more documents at the top of the retrieval results based on different criteria (Si and Callan, 2003, 2004; Shokouhi, 2007; Seo and Croft, 2008; Paltoglou et al., 2008; Ipeirotis and Gravano, 2008; Thomas and Shokouhi, 2009; Kulkarni et al., 2012; Markov and Crestani, 2014). The uncooperative environment depends on sampling and centralised sample index technique is out of scope in this thesis and I just study the cooperative environments.

The effective state-of-the-art routing approaches are classification-based methods (Xu and Li, 2007a; Arguello et al., 2009a; Cetintas et al., 2009; Hong et al.,

2010). These approaches use current machine learning classification techniques to build a classifier over a training set of features and labels in order to predict the scores of resources for a future query (or a testing query). The classification-based approaches, train a classifier model on specific features, obtain advantages which include (Arguello et al., 2009a): (i) the flexibility of embedding many features; (ii) easy offline generating training set; (iii) the choice of using different machine learning algorithms; (iv) and the effectiveness over the current resource selection approaches such as Taily (Aly et al., 2013) and Relevant Document Distribution Estimation (ReDDDE) (Si and Callan, 2003) in automatically tuning parameters, combining multiple evidence, and avoiding over-fitting.

Users on the web should collectively provide a search service and have a full control over what information they wish to share as well as how they share it. This phenomenon can be achieved through using Peer-to-Peer Information Retrieval (P2P-IR) systems, in contrast to the other forms of federated search (i.e, meta-search and vertical search). These forms are static resources of search engines and verticals, in which the users do not interfere in uploading their own information or even give any services. P2P-IR systems, as an application of P2P networks, have a number of similarities with that of federated information systems in search processes (Lu, 2007; Klampanos and Jose, 2012; Tigelaar et al., 2012). The search process in P2P networks is somewhat complicated in peers' representations, resource selection (or query routing), and result merging (or result fusing).

However, like in federated information systems, there are three major challenges in P2P-IR systems, which have to be taken into account to design an effective and efficient retrieval system. These challenges are complicated in comparison with vertical and meta-search systems in acquiring the resources information, query routing and result merging at multiple points in the networks. The challenges in P2P-IR networks include (Callan, 2000; Tigelaar et al., 2012; Klampanos and Jose, 2012): (i) resource representation that represents the useful information of the distributed peers and *local search directories* ; (ii) resource selection that selects and routes a given query to the relevant peers and/or *service directories (or super-peers)*; (iii) and results merging that combines the search results from multiple resources including super-peers into a final result list. The difficulty in using merging process in P2P networks occurs in incorporating the

results of multiple peers at different points in the network. However, resource selection (i.e. query routing) is a critical component in P2P-IR; low-quality resource (peer) selection, the case where peers contain relevant documents get excluded while executing the query routing, would inevitably lead to less effective retrieval results. One of the difficulties in P2P networks is that it is almost impossible to collect global statistics, which are needed to be estimated to route queries to those relevant peers (Richardson and Cox, 2014). Flooding approach does not require global statistical information to route queries across the network as they are routed to the whole resources, however, results in generating a number of messages across the network and subsequently affecting the limited bandwidth. This approach also increases the response time for the query and overwhelms the final result list with more noisy (non-relevant) documents.

This chapter overviews the resource selection techniques in cooperative federated search environments including classification-based resource selection approaches. In addition, a survey of P2P networks is given, including its concept, architecture, topologies, P2P-IR applications, some related work on query routing for information retrieval in these networks, and reputation-based systems in P2P networks.

3.2 Cooperative Resource Selection Techniques

Recent cooperative resource selection techniques assume that any resource is a big document of small documents in the resource's collection (Shokouhi and Si, 2011; Crestani and Markov, 2013). In this case, the broker builds its decision on selecting specific resources by deploying the traditional retrieval models at a big document collection level. In particular, the selection algorithms obtain detailed lexical statistics of the connected resources in cooperative environments. The relevance of a resource, therefore, depends on estimating the lexical similarity of that resources to the given query. I discuss some of these techniques as follows:

GlOSS. Gravano et al. (1994) proposed an initiated version of GlOSS (Glossary-Of-Servers Server) as a resource selection for Boolean IR retrieval, also known as bGlOSS. The bGlOSS approach supports Boolean queries and just requires the resource size values and query term frequency information. The number of doc-

uments that have all the query terms is estimated as follows:

$$bGIOSS(Q, R_i) = \frac{\prod_{t \in Q} tf_{t,R_i}}{|R|^{|Q|-1}} \tag{3.1}$$

Resources are ranked based on their estimated number of answers for the given query. Later, GlOSS was generalized to gGlOSS (or vGlOSS) (Gravano and Garcia-Molina, 1995; Gravano et al., 1999) as a vector space model version to be used by IR models. In gGlOSS, resources are selected based on their goodness values with respect to any particular query defined as:

$$Goodness(q, l, R) = \sum_{d \in \{R | sim(q,d) > l\}} sim(q, d) \tag{3.2}$$

Where $sim(q, d)$ is a function that calculates the cosine similarity between query q and document d (Salton and McGill, 1986; Salton et al., 1983). The summation of the vectors of all documents' vectors in a specific resource with a given query vector determines the goodness of the resource to that query. The threshold l is used to reduce the noise values in low similarity. Therefore, once $Goodness(q, l, R)$ is calculated for each resource R with respect to query q at threshold l, the ideal rank for the query at threshold l can be constituted by sorting the resources in descending order of their goodness.

CORI. COllection **R**etrieval **I**nference **net**work (CORI net) (Callan et al., 1995; Callan, 2000) is one of the most popular selection technique that depends on the Bayesian inference network model with an adapted Okapi term frequency normalization (Robertson and Walker, 1994). In particular, the CORI approach measures belief values for individual resources using the Bayesian inference network model. The evaluation effectiveness of CORI technique was implemented on INQUERY ad-hoc retrieval system (Turtle, 1991; Turtle and Croft, 1990). The belief value of i^{th} resource related to word t, is calculated as follows:

$$Score_{R_i} = P(Q/R_i) = \frac{1}{|Q|} \sum_{t \in Q} b + (1 - b) \times T_i \times I_i \tag{3.3}$$

$|Q|$ is the number of terms in the query, T_i is the tf analogous for term i, that

is:

$$T_i = \frac{df_{t,i}}{df_{t,i} + 50 + 150 \times rw_i / avg\_rw} \tag{3.4}$$

and I_i is the idf analogue for term i, that is:

$$I_i = \frac{log(\frac{N_r + 0.5}{cf_t})}{log(N_r + 1.0)} \tag{3.5}$$

where $df_{t,i}$ (documents frequency) is the number of documents in the i^{th} resource that contain t; cf_t (collections frequency) is the number of resources that contain t; N_r is the total number of available resources. rw_i is the total number of words in the i^{th} resource, and avg_rw is the average rw of all resources. Finally, b value is the default belief that refers to the minimum term frequency component, which is usually set to 0.4 in earlier experiments (Callan et al., 1995). Therefore, $P(Q/R_i)$ is used by the CORI algorithm to rank resources, which mimics Okapi-BM25 retrieval model in ranking documents as calculated in Equation 2.16.

CVV. Cue-**V**alidity **V**ariance (Yuwono and Lee, 1997) was used as a resource selection technique in the WISE index server (Yuwono and Lee, 1996). The CVV algorithm uses only document frequency information of resources and calculates the goodness of a specific resource R_i for n-terms query q as follows:

$$Score_{R_i} = Goodness(R_i, q) = \sum_{t \in Q} CVV_i . df_{t,i} \tag{3.6}$$

where $df_{t,i}$ refers to the document frequency of query term t in resource R_i and CVV_i is the variance of cue-validity (CV_i) of that term (Goldberg, 1995). The value of $CV_{R_i,t}$ gives an indication of how close is the t term in the query to resource R_i from other resources. In other words, the CVV component estimates whether a term is useful for differing one resource from another, which is calculated as follows:

$$CV_{R_j,t} = \frac{\frac{df_{t,R_j}}{|R_j|}}{\frac{df_{t,R_j}}{|R_j|} + \frac{\sum_{k \neq j}^{N_R} df_{t,R_k}}{\sum_{k \neq j}^{N_R} |R_k|}} \tag{3.7}$$

In such equation $|R_k|$ is the number of documents in resource R_k and N_R is

the total number of resources. The variance of cue-validity CVV_i is calculated as:

$$CVV_t = \frac{\sum_{j=1}^{N_R}(CV_{R_j,t} - \overline{CV_t})^2}{N_R} \tag{3.8}$$

Here, $\overline{CV_t}$ represents the average $CV_{R_j,i}$ for whole resources which is defined as follows:

$$\overline{CV_t} = \frac{\sum_{j=1}^{N_R} CV_{R_j,t}}{N_R} \tag{3.9}$$

Kullback-Leibler (**KL**) divergence resource selection was proposed by Xu and Croft (1999) that depends on document clustering and language modeling. The resources' representations and user queries are treated as multinomial distributions. The k-means clustering algorithm (Jain and Dubes, 1988) was used for grouping documents based on topics and then the Kullback-Leibler divergence (Lafferty and Zhai, 2001) was used between the distribution of the query and the resources' topics to measure how likely the available resources' descriptions matching the query for routing processes. The distance between a query Q and a resource R is estimated as follows:

$$P(Q/R) = \sum_{t \in Q} \frac{f(Q,t)}{|Q|} log \frac{\frac{f(Q,t)}{|Q|}}{\frac{(f(R,t)+f(Q,t))}{(|Q|+|R|)}} \tag{3.10}$$

Here, $f(Q,t)$ is the number of occurrencess of term t in the query, $|Q|$ is the number of terms occurrences in the query, $f(R,t)$ is the number of occurrence of the term t in the resource, and $|R|$ is the total number of terms occurrence in the resource R.

LM. Si et al. (2002) proposed a language model framework for resource selection and result merging. The resource selection part of the framework depends on the query-based sampling approach to generate the resources' representations (Callan and Connell, 2001; Callan et al., 1999). The generated resources' representations were used to build a language model for each resource from its documents and a global language model from all sampled documents of all the resources. To find the largest probabilities of query and resources, the system ranks resources by measuring the Kullback-Leibler divergence (Lafferty and Zhai, 2001)

between the query model and the resources' models.

$$P(Q/R) = \prod_{q \in Q} (\lambda P(q/R)) + (1 - \lambda) P(q/G) \tag{3.11}$$

Where $P(Q/R)$ is the language model for the resource R and $P(Q/G)$ is the language model for all resources, which overestimates the document-query similarity. The linear interpolation constant λ smoothes the resource-based language model with the global language model and is usually adjusted in the range of 0 to 1. Resources with the largest probabilities $P(Q/R)$ will be selected as they are more likely to have more relevant documents. Such resource selection method is similar to the method in (Xu and Croft, 1999) where the equation is simply taking the *log* in equation 3.11 as follows:

$$KL(Q/R) = \sum_{q \in Q} P(q/Q) log(\frac{P(q/Q)}{\lambda P(q_i/R) + (1 - \lambda) P(q_i/G)}) \tag{3.12}$$

Where resources are ranked using the negative of the KL divergence. Both methods are using the word distribution as a basis for the similarity measurement.

PRF. Pseudo-**R**elevance **F**eedback selection algorithm is proposed by Chernov et al. (2007), that is deployed in the Minerva project, which is a structured P2P web search engine (Bender et al., 2004). In Minerva, all peers cooperatively maintain peer-summary information about which peer has documents for which index terms. Such information is organized in peer lists constructed for each term in the system. For sharing the peer-summary information, Minerva disseminates all peer lists using Chord protocol (Stoica et al., 2003) where each term is hashed to peer network address to know which peer is responsible for managing which peer list. However, the PRF method consists of two steps for selecting most relevant resources for a given query. Firstly, the method selects a set of relevant resources (peers) using the language selection approach in (Si et al., 2002) where the global language model is measured by an approximation of peer lists related to a specific query, which is defined as follows:

$$Score(Q, R_i) = - \sum_{t_j \in Q} log P(t_j/R_i) \tag{3.13}$$

$$P(t_j/R_i) = \lambda.\frac{rtf_{t_j}}{|R_i|} + (1-\lambda).P(t_j/G) \tag{3.14}$$

$$P(t_j/G) = \frac{\sum_{i=1}^{|R|} rtf_{t_j}^{R_i}}{\sum_{j=1}^{|T|}\sum_{i=1}^{|R|} rtf_{t_j}^{R_i}} \tag{3.15}$$

where $P(t_j/R_i)$ is the probability of term t_j of the language model for resource R_i; λ is an empirically set smoothing parameter between 0 and 1; rtf_{t_j} is the resource term frequency (the number of term occurrence in the resource); T is a system vocabulary (the full set of distinct terms on all resources); $P(t_j/G)$ is the generation probability of term t_j of global language model for all resources G.

Secondly, the relevant resources from the first step is used to retrieve relevant documents by executing a query and then the top-k ranked documents from the relevant resources are used by ad-hoc query expansion techniques (Ponte and Croft, 1998a; Robertson and Walker, 1999) to extend the query terms. Also, the top-k ranked documents are used by query modelling techniques (Tao and Zhai, 2004; Zhai and Lafferty, 2001) to estimate the generation probability $P(t_j/PR)$ for each query term from the pseudo-relevance language model PR. A cross-entropy, an information-theoretic measure of distribution between two distributions, is used to select the most relevant resources as follows:

$$H(Q, R_i) = -\sum_{t_i \in Q} P(t_j/PR).logP(t_j/R_i). \tag{3.16}$$

where the lower cross-entropy of a resource language model (the pseudo-relevance based language model), the higher similarity of these models. The two values expressing the importance of a query term: query-specific $P(t_j/PR)$ and global language model $P(t_j/G)$.

Melucci and Poggiani (2007) proposed a resource selection method as a weighting scheme adapted from retrieval models for IR across P2P networks (**TF-IDF**). The method selects super-peers, peers, and documents at three levels of selection to improve the retrieval effectiveness. The results show that after contacting about 16% of the peers, a system based on the proposed weighting scheme can retrieve about 40% of the relevant documents that can be retrieved by a centralized system.

PCAP (Puppin et al., 2010) exploits query logs for improved resource selection. A query log is used to build a matrix where each document-query combination is assigned relevance scores. This matrix is then co-clustered to identify clusters that have two parts, a set of documents and a set of queries for which the documents are relevant. These separate co-clusters are then managed by separated peers, and a subset of co-clusters are chosen to route each query. The focus here is to find resource selection and the clustering is done on query logs, and thus, PCAP cannot work in the absence of accumulated historical query log information.

Taily (Aly et al., 2013) is a large document-based resource selection method proposed for shard environments. The shard is selected depending on query independent features that are modelled as Gamma distribution extracted offline from the whole collection and shards. The high shards' documents' scores on the right tail of the distribution determine the selection score of the shard. The experimentations show that the Taily approach obtains competitive effectiveness and efficiency results in comparison with baseline methods. The problem in using the Taily approach in P2P systems is to gather the whole collections' statistical information that is scattered remotely in different peers.

Cooperative resource selection techniques require statistical information about resources for query routing decision-making. Each method requires different statistical information about the query terms from resources and selects them based on its formulation of ranking. These techniques were studied in various environments such as meta-search and P2P networks. In this thesis, I select the effective and prominent cooperative methods which are GIOSS, CORI, CVV, KL, TF-IDF, and Taily and leave the others for future work. I use and study the effectiveness of these techniques in cooperative federated search environments such as meta-search and the target semi-structured P2P-IR model (i.e, K-means model).

3.3 Classification-based Resource Selection Approaches

In Section 3.2, I explained the standard cooperative resource selection methods that depend on the query terms' statistical information to select the relevant

resources. The standard resource selection methods need to manually tune their parameters to make a decision on selecting resources. This section surveys prior research on the state-of-the-art classification-based resource selection methods that select resources based on machine learning classifier model built on training set of past queries. The classification-based approaches has a set of advantages over standard resource selection methods as discussed in Introduction 3.1.

Arguello et al. (2009a) proposed a logistic regression learner for each resource to route a new query. The training set is generated offline using three categories of features and full-data set retrieval to assign labels. The features include (i) corpus-based features such as CORI (Callan et al., 1995), a variant of uncooperative ReDDE approach called ReDDE.top (Arguello et al., 2009b), and an uncooperative geometric average approach (GAVG) (Seo and Croft, 2008), (ii) query-categorical features such as query topic, (iii) and click-through features which are a signal of collection relevance derived from queries' clicks on its documents. However, a centralised full-dataset retrieval is used to generate a label for each query-resource pair where the number of documents that are retrieved for a query on the centralised full-dataset and they are in the collection's (resource's) documents is considered as a label for that resource. The results show the same level or in some cases significantly better effectiveness than all single corpus-based features as baseline methods.

Cetintas et al. (2009) proposed a resource selection method that utilises the past queries' results to rank and route a query to the most relevant resources. The proposed approach is motivated as in real scenarios users'queries are similar and duplicates. The relevance of each resource on past queries is calculated after running a specific resource selection method and testing the merged result list at a specific threshold. This method uses a regression model for result merging that based on training downloaded documents of the past queries which also used to estimate the similarity between past query and the current testing query on common documents' scores. Consequently, in aggregating the two steps above, the relevance score of a specific resource is a summation of past queries' scores on that resource multiply by the similarity of past queries and the current testing query.

Hong et al. (2010) proposed a novel joint probability classification approach for resource selection. This approach is built based on an assumption of that information resource similar to other relevant resources has a high probability to be relevant. In particular, the approach builds a logistic model as an independent model to combine all the features of individual resources in the training step. These features are CORI (Callan et al., 1995), geometric average (GAVG) (Seo and Croft, 2008), modified ReDDE that is called ReDDE.top (Arguello et al., 2009b), and a score estimated on top-ranked merged result list that is retrieved from centralised index of sample documents. Then, as a joint classification probability score, the logistic model is aggregated with a similarity term score modelled between two sources. However, a set of similarity metrics were evaluated that depend on the effectiveness of resources on training queries and the Kullback Leibler method between the language models of the two resource contents. Intensive experiments show the consistent effectiveness of the proposed method on different testbeds based on independent and joint probability classification models.

Xu and Li (2007a) proposed new features trained in a pointwise learning to rank algorithm. The adopted learning functions are RankSVM (Joachims, 2002) and the original SVM methods that trained over 20 query-independent and query-dependent features along with human assessments to assign a label for each collection. The RankSVM approach is an ordinal regression-based algorithm that has five categories of labels, while SVM is a classification-based algorithm with positive and negative labels. The experimental results show the effectiveness of learner algorithms in comparison with the CORI resource selection algorithm as well as the benefits of the suggested features in resource selection.

Classification-based resource selection methods use classification machine learning algorithms to create a classifier based on a training set of specific features and labels on all resources for ranking resources. The discussed classification-based techniques use different features and labels on specific environment and showed effective results using such methods. In this thesis, I use these techniques in meta-search and the target semi-structured P2P-IR model using a machine learning classifier that is built up on well-studied training set of features and labels.

3.4 Peer to Peer Networks

3.4.1 Peer to Peer Concept and Paradigm

Peer to Peer (P2P) technology provides a network paradigm of a set of distributed and participated nodes (also called peers) logically connected through a protocol layer (Androutsellis-Theotokis and Spinellis, 2004; Klampanos and Jose, 2012; Tigelaar et al., 2012; Lu, 2007; Lua et al., 2005). The peers are self-organised and logically connected to an overlay layer that is not necessarily associated with the underlying physical connections as shown in Figure 3.2.

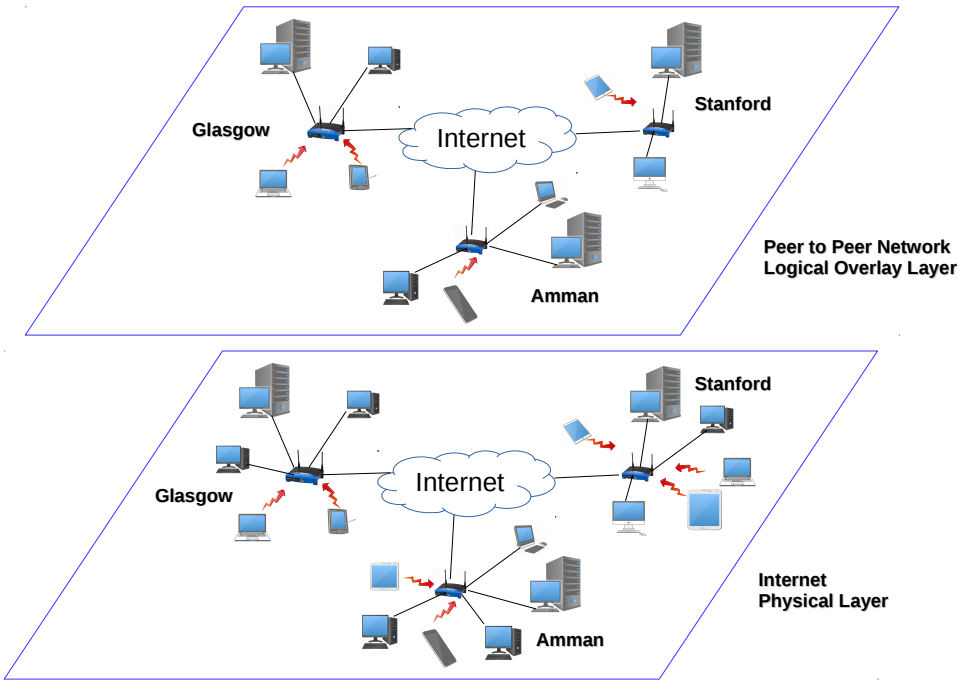


Figure 3.2: P2P overlay Network

In spite of its simplicity, there are many definitions related to P2P overlay network concept, which can be inferred as: *P2P overlay network consists of a set of* ***self-organising*** *and distributed* ***peers*** *that communicate among each other in a* ***logical decentralised and symmetric manner*** *over the physical layer of the Internet network. The quality of peers is defined by their level of* ***willingness in participation*** *to share resources such as CPU cycle, storage, contents, and bandwidth that will be added to a whole available capacity.*

As an example shown in Figure 3.2, I select three networks from different locations in the world; Amman (Asia, Jordan), Stanford (USA, America), and Glasgow (Europe, Scotland-UK). The peers, in overlay layer, can communicate with each other without centralised control. A few peers from the networks of three cities do not participate in the P2P overlay network, although they are physically connected. Under such a logical communication, the participating peers install a client application to initiate their activities in P2P network in order to provide or request services. However, P2P overlay networks can deploy various applications to supply services for the users who sit behind such peers like content distribution, file-sharing, instance messaging, streaming media, telephony, and Information Retrieval (IR), etc. (Tigelaar et al., 2012). In this thesis, I focus on the information retrieval application, considering the peers in P2P-IR overlay network are categorised into three component units with a possible combination of them. The component units are an information provider, consumer (request information), and service provider (also called super-peers or Hubs), in which it facilitates the effectiveness and the efficiency of searching relevant information (Klampanos and Jose, 2007; Lu and Callan, 2003).

3.4.2 Peer to Peer Architectures

The network architecture of P2P networks determines the functionality and responsibility of each peer based on its willingness. Based on the P2P overlay network definition, P2P networks can be classified into two basic architectures: structured and unstructured networks, which are clarified as following subsections.

3.4.2.1 Unstructured P2P networks

In unstructured P2P networks (Lu and Callan, 2003; Klampanos and Jose, 2004; Yang and Fei, 2009; Chen et al., 2012), the location of documents is completely separated from the network topology, where each peer has their own contents and has full control in organising them. Users, in the unstructured network, can provide complex queries and randomly join the system without any prior knowledge of the network topology. Unstructured networks can be classified into three types based on the degree of centralisation by taking peers' willingness

(or participation) into account, which are centralised P2P architecture; purely decentralised P2P architecture; and hierarchical P2P architecture (Androutsellis-Theotokis and Spinellis, 2004) as shown in Figure 3.3.

Centralised P2P architecture

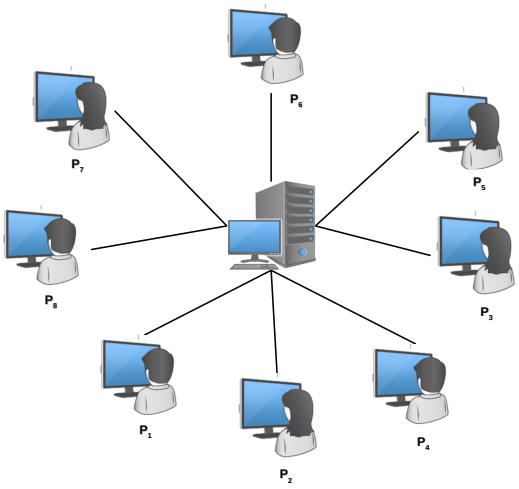
Centralised P2P architecture has a single centralised server (or broker) as logical directory services that stores meta-data (or description) information of the other peers' content to facilitate the search process. Each broker has strong computing capabilities and storage. The consumer peers send a query to the broker in order to locate information provider peers that contain the relevant content for the given query. Then, the broker sends the contact information of the relevant information provider peers (e.g., IP address and port) back to the requested and consumer peer to create a direct connection for retrieving the relevant documents. Centralised P2P architectures discover resources for peers in a very simple, flexible, and efficient manner, although a single point of failure might expose to collapse the whole system or be a hotspot (bottleneck) occurred from the peers during search processes. The most famous architecture of this network is the Napster music file-sharing system[1].

Purely decentralised architecture

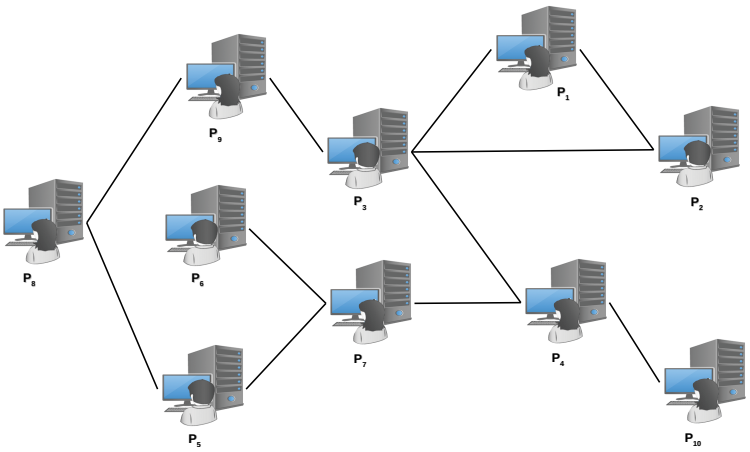
In purely decentralised networks (or flat P2P architecture), the peers, in the network, are totally decentralised and act simultaneously as an information provider, consumer, and directory service (or SERVENT (SERVer and cliENT)) with the same level of participation in sharing content. Each peer stores its own content and sends its descriptions to the neighbouring peers. In the search process, a consumer sends a query to all neighbouring peers. Then, the neighbouring peers process the query and forward it to their neighbours as a consecutive chain between peers to find specific information. This search process is stopped when the required data has been found or the predefined Time To Life (TTL)[2] is reached. A purely decentralised architecture uses two search mechanisms, which are flood-

[1] www.napster.com (October, 2016)

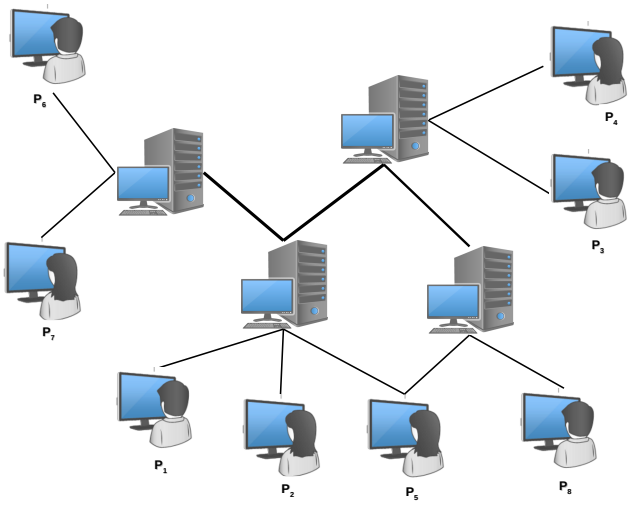
[2] TTL is a limit on the number of steps for a query in the network before it expires through decreasing it at each step (Lv et al., 2002)



(a) Centralised P2P Architecture.



(b) Purely Decentralised P2P Architecture.



(c) Hierarchial P2P Architecture.

Figure 3.3: Unstructured P2P Architectures

ing and random walk, depends on forwarding the query to the whole or random neighbours to reduce network traffic cost. Although purely decentralised networks are very simple in joining and leaving peers, low cost of maintenance, and fault-tolerance towards peers or network failures, they have many weaknesses such as: (i) they are poor in locating rare items; (ii) they are not scalable as the load on each peer increased with the total number of queries and the system size, (iii) and they consume high bandwidths and resources. Gnutella v0.4[1] is considered as one of the most famous examples of purely decentralised P2P networks.

Hierarchical P2P architecture

Hierarchical decentralised architectures (known as super-peer architectures) contain two types of peers; super and regular peers. The super-peers (or hubs) have a high level of willingness to store the meta-data of their regular peers and other super-peers. The super-peers also create a connection between each other to submit and answer queries on behalf of their own regular peers (Yang and Garcia-Molina, 2002). To find relevant documents, a peer sends a query to its connected super-peer(s), which routes a query to its peers and other super-peers. The super-peers route a query to their peers and merge the retrieved result lists to be sent back to the requesting super-peer. The requesting super-peer merges the super-peers' result lists and sends back the final merged result list to the requesting peer (or consumer). Hierarchical decentralised systems combine the advantages of the other two centralised and purely decentralised systems in load balancing between super and regular peers, through providing heterogeneity across peers, to improve the performance (Yang and Garcia-Molina, 2003). The disadvantage appears in failures of a few super-peers that will have an impact on the system. The most common examples of this architecture are a modified version of Gnutella 4.0 called Gnutella 0.6[2], KaZaA[3] and JXTA[4].

[1] https://courses.cs.washington.edu/courses/cse522/05au/gnutella_protocol_0.4.pdf, (October, 2016)

[2] http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html, (October, 2016)

[3] www.kazaa.com (October, 2016)

[4] http://www.oracle.com/technetwork/java/index-jsp-136561.html, (October, 2016)

3.4.2.2 Structured P2P networks

Structured networks are tightly controlled where the location of content is out of the peers' control (Stoica et al., 2003; Ratnasamy et al., 2001; Maymounkov and Mazières, 2002). In particular, a new peer joins the system with predefined rules in a specific location using a data management protocol. A structured P2P network uses a Distributed Hash Table (DHT) to provide a map between content (file identifier) and its location over peers (peer IP address). The DHT table is distributed over peers to store and retrieve relevant content in the network. Hence, the peers act as SERVENT with equal functionality as in purely decentralised P2P networks. Although structured networks provide efficient discovery of data items using the given query key, this system does not support complex queries where the query must have an exact matching name (key) with stored data object to retrieve the same peer of the data object.

3.4.3 Peer to Peer Topology

Network topology, in P2P network architecture, determines the organisation of peers over a logical layer of the network. The location of a peer on the overlay layer protocol has an effect in navigating the relevant peers to evaluate the given query. A network topology of peers and links between them exhibits three concepts of peer distance which are graph distance, content distance, and latency distance (Lu, 2007). Graph distance refers to the (shortest) path between two peers, while content distance and latency distance in a network architecture refer to peer distance in content and latency response time distance, respectively. These concepts of peer distance are used to enhance the effectiveness and efficiency of search in three network topologies of P2P networks, which are interest-based locality, content-based locality, and small-world networks. A P2P structured (or unstructured) architecture might have a combination of the three properties in a non-mutually exclusive manner.

3.4.3.1 Interest-based Locality

A P2P network topology with interest-based locality aggregates the peers with a similar interest in content into the same location near to each other with short-

est path distances. This scheme clarifies the relationship between a peer interest based on its past queries and the other peer's content to reduce the network traffic for future requests (Krishnamurthy and Wang, 2000). Ramanathan et al. (2002) proposed a P2P interest-based network topology where the peers are clustered based on the frequency of responses of the past queries to enhance query routing of similar interest peers for future queries. Sripanidkulchai and Zhang (2005) built a loosely interest-based shortcut topology over Gnutella as purely decentralised P2P network based on past queries response. Leveraging the highest probability of past queries response, Shao and Wang (2005) constructed a BuddyNet topology over a hierarchical P2P network to organise the peers around their friend peers of the same interest. Yeferny et al. (2013) proposed a semantic cluster network topology that aggregates the peers having similar interest learned from past queries into the same cluster. A peer searches for similar friends' interests through forwarding its interest vectors to establish a connection with those who are similar in interests and share its knowledge interest with them.

Although these research works use past queries as an indication of a peer interest, they suffer from multiple peer's interests that might bias the interest of peers to a far distance in the network.

3.4.3.2 Content-based Locality

Content-based locality scheme, in P2P networks, organises the peers based on content similarity within a shortest path area (Klampanos and Jose, 2007; Lu, 2007). This leads to the assumption of the cluster hypothesis that states to group similar documents of the same request around each other (van Rijsbergen, 1979). In P2P networks, the hypothesis ensures organising the relevant documents of a specific query into semantically homogeneous peers. Therefore, peer distance depends on content distance in organising the network topology. Content-based locality schemes help in routing a query to a small number of peers that have content topically similar to the given query. This scheme demonstrates the importance of clustering in organising the peers in P2P networks into a topical homogeneous network for effective retrieval and efficient message routing as will be discussed in more detail in Subsection 3.4.4.

3.4.3.3 Small World

Small world networks are another types of P2P network topologies used for organising peers (Watts and Strogatz, 1998; Kleinberg, 2000). The idea of small world network, as a promising evolution, is established after the theory of six degrees of separation between two people in the world (Goth, 2012). However, small world networks form a middle ground between regular and random networks, where any two nodes in the network are likely to be connected within a shortest path through a sequence of intermediate nodes (Kleinberg, 2000). Hence, small world networks are commonly pervasive in large-scale sparse networks, including a well-known hierarchical file-sharing P2P network Gnutella v0.6 (Stutzbach and Rejaie, 2005). The small world network has two properties, which are short global separation as regular networks and high local clustering of nodes as random networks (Watts and Strogatz, 1998). The interest-based and content-based locality could be seen as a small world network if the clustering of peers is conducted on past queries' interests or similarity of content topics.

3.4.4 Peer to Peer Information Retrieval

Peer to Peer Information Retrieval (P2P-IR) systems are considered as one of the federated search forms with more complexity in routing and merging between participating peers (Lu, 2007; Klampanos and Jose, 2012; Tigelaar et al., 2012). In this Subsection, I discuss the challenges in building an effective P2P-IR system, P2P-IR environments, cooperative P2P-IR environment including peer representation, query routing and result merging components, and some related work on query routing and retrieval on P2P networks.

3.4.4.1 P2P-IR Challenges

The research works on P2P information retrieval system have resulted in poor performance and as a result, initial enthusiasms have been receded (Lu, 2007; Klampanos and Jose, 2007; Cuenca-Acuna et al., 2003; Chernov et al., 2007; Koskela et al., 2013). Tigelaar et al. (2012) presented the main technical challenges in developing an effective P2P-IR, which includes: (i) latency in query response time that results from a number of messages sent across the network

and ideally should be competitive with centralised search response time. Latency increases with the number of routed peers in the system that inevitably affects user satisfaction. Therefore, efficient query routing is needed to reduce flooding the network with messages, which in turn has an adverse effect on the system effectiveness; (ii) a dynamic document churn rate in the system that has an effect on updating the distributed indices and creates difficulties in P2P architectures to collect global statistics for routing query to most relevant peers (Richardson and Cox, 2014); and (iii) a lack of standard testbeds for P2P-IR evaluation that cover the P2P real-life scenarios such as churn rate of peers, dynamic document distribution, and simulating user behaviour. Although Klampanos et. al. (Klampanos et al., 2005) suggested a number of P2P-IR testbeds based on real-life scenarios, the evaluation is still a neglected task in P2P-IR systems.

3.4.4.2 P2P-IR Environments

The P2P-IR systems have two environments as a federated search form, which are cooperative and uncooperative environments. In cooperative P2P-IR environments, each document provider in the network can cooperate closely to inform their neighbours about their contents by providing useful information such as document copies or content statistics for document retrieval and ranking (Chen et al., 2009; Klampanos and Jose, 2007; Puh et al., 2008). The cooperative P2P-IR networks have three characteristics (Zhang, 2011): (i) each document provider needs to publish full document copies or document descriptions (term statistics) to the network, (ii) content descriptions are not restricted and can be used by other peers; and (iii) the same search engine should be employed by every peer. On the other hand, in uncooperative environments, a document provider may not be willing to provide their individual document copies or descriptions to the network due to copyright issues and access limitations such as digital library search (Markov and Crestani, 2014).

3.4.4.3 Cooperative P2P-IR Environment

The P2P-IR paradigm has been investigated for various P2P network topologies. In this thesis, I use the cooperative semi-structured P2P-IR environments. The process of search across cooperative P2P-IR networks requires each peer

to provide a statistical information (or peer description) about its corpus. The statistical information is critical to accurately identify the interest of peers in sharing the content. However, the peers' presentations vary in different kind of provided information, which include (Lu, 2007): (i) named-based representation; each peer gives the terms included in its documents' titles, (ii) free-text representation; including a set of keywords from the documents of a peer or is assigned manually by the peer, (iii) controlled-vocabulary representation; a set of terms that is generated from a controlled vocabulary based on an ontology, (iv) and full-text representation; this kind of information can be categorised into two representations, which are full-terms and cluster-terms. The full-text representation includes all the terms from the documents of a peer along with their weights. The cluster-terms representation includes the terms of cluster centroids that are generated by running a clustering algorithm on the peers' documents or might be aggregated across peers' centroids along with their weights (Klampanos and Jose, 2007; Alkhawaldeh and Jose, 2015). On the other hand, resource ranking and selection (or query routing) is one of the essential challenging problems in P2P-IR systems. This is due to routing a query to non-relevant peers inevitably results in low-quality retrieval effectiveness. Hence there have been a set of techniques to rank the peers that are most likely to contain relevant documents. The selection decision depends on how accurate is the peer's description (or resources' statistics) to a given query. However, in addition, P2P-IR systems can use the cooperative resource selection techniques discussed in Sections 3.2 and 3.3. The third challenge in P2P-IR networks is the result merging (or result fusing) that combines the result lists of requested peers at different points in the network along with the query path to the query sender.

3.4.4.4 Related Work on P2P-IR Systems

In this thesis, I study query routing on cooperative environments of semi-structured P2P-IR networks. Research works, related to this work, are explained to highlight the importance of these works in improving the query routing quality. Table 3.1 briefly summarises the related work, while I discuss them in more detail below.

- Yang and Garcia-Molina (2003) conducted an experimental study of super-peer performance, in a P2P network, due to build robust super-peer net-

Table 3.1: Related work on P2P-IR Systems

	Architecture			Locality			Resource Representation				Query Routing		
Method	Structured	Decentalised	Hierarchial	Interest	Content	Small-world	Named-based	Ontology	Full-terms	Cluster-terms	Flooding	Random Walk	Selection
PlanetP		✓			✓				✓				Gossiping
Lu			✓		✓	✓	✓		✓		✓	✓	KL
Jin		✓				✓			✓				Semantic
Klampanos			✓		✓	✓				✓	✓		Semantic
iCluster		✓			✓	✓				✓			Semantic
SOON		✓				✓		✓					Semantic
Trust-aware	✓								✓				KL + Structured
Class-based		✓			✓					✓		✓	Semantic
BloomCast			✓						✓			✓	LM
Paradox		✓										✓	

works and to explain the amount of nodes' work per unit time and the number of the result returned. A set of parameter settings were determined such as the topology of the network and user behaviour over the file sharing application system. The topological parameters include graph settings of graph type, size, cluster size and number of nodes as well as TTL values. The user behaviour consists of query formulation over time by the users. In addition, clustering mechanisms were utilised for organising peers in a group whereby each group act as similar content (topics) and creating a hierarchical structure on top of a flat network.

- PlanetP (Cuenca-Acuna et al., 2003) is a P2P information retrieval network where the peers share a summary description of their local collection as term vectors encoded in bloom filters. The encoded term vectors disseminated over the network to rest of peers using a gossiping algorithm. The routing process occurs after ranking the peers by calculating the closeness of their local descriptions to a given query using Inverse Peer Frequency metric (IPF_t). Then, the initiated peer sends the query to the top-ranked peers and merges their results as a final list. The evaluation of such approaches was limited to structured retrieval and suffers from scalability issues.

- Lu and Callan (2003) proposed a content-based resource selection algorithm in a hybrid P2P-IR network and evaluated it using a large-scale digital library testbed. In this network, the super-peers use an adapted K-L divergence-based resource selection algorithm (**KL**) (Xu and Croft, 1999) to rank their leaf peers and other super-peers based on their content representation for routing the given query. They compared the proposed

algorithm to two baseline methods; the name-based model that uses hashed terms of documents' title names in the peers and matching-based model that uses peers' vocabularies for query matching process. The results showed a high accuracy and efficiency in the proposed model to the two baseline algorithms. An extensive experimental study on adaptive resource selection and result merging algorithms was conducted on the same testbed in a hybrid P2P network (Lu and Callan, 2005; Lu, 2007), and the results were compared to flooding and random walk algorithms. The experiments were run on cooperative and uncooperative environments of hybrid P2P architecture. The results showed also more efficient and accuracy in retrieval performance as in their previous work. Even though improvements are shown with respect to the baselines, the retrieval effectiveness is very poor in comparison to the centralised search systems. In addition, the user behaviour, including long-term interest-based on past queries, is exploited in full-text federated search in a P2P network to build a model to improve the efficiency of future queries of similar interest (Lu and Callan, 2006). In particular, the past queries and ranked search results are used to build an efficient model at the super-peer level to evaluate the query routing performance at that level for future queries.

- Jin et al. (2006) proposed a semantic search approach in an unstructured P2P small-world network. The search process is conducted by sending a query along with its categorised topic to the most topical semantic peers. The sender peers select the most semantic similarity peers to answer the given query. The semantic similarity of a peer is estimated on the number of documents on this peer and the number of documents within the same query topic. The experimental results show better performance in efficiency and Recall rate as retrieval effectiveness over Gnutella 4.0 and the interest-based shortcut model of 150% and 60%, respectively. The method suffers from aggregating the required information to compute the semantic similarity such as the size of the peers and its documents' topics.

- Klampanos and Jose (2003, 2004, 2007) proposed a cluster-based 2-tier P2P architecture. The idea is to identify the content categories of peers using in-

peer documents clustering by using an agglomerative method called ward clustering and then group them using single-pass clustering algorithm at the super-peer level. Although, the approach is evaluated on a suite of large-scale testbeds for realistic evaluation (Klampanos et al., 2005), the retrieval effectiveness of the system is not satisfactory. Applying the single-pass clustering, to identify the semantic categories at the super-peer level for a large collection of documents, is computationally expensive and also it depends on the threshold used (Klampanos et al., 2006). In addition, the poor results of the architecture occurred due to the noise of the clustering method on web documents, which affects the retrieval results and query routing. Hence, computationally feasible and alternative models need to be explored. They further proposed two features for improving the retrieval performance include a replication of relevant documents over the network and using relevance feedback to increase the values of past queries' terms at the super-peer level.

- iCluster (Raftopoulou and Petrakis, 2008; Raftopoulou et al., 2008) is a periodic rewiring approach organises the peers in a P2P network into clusters of similar interests to improve the retrieval effectiveness and network load routing efficiency. Each peer groups its documents using a clustering algorithm to form a set of interest centroids where each centroid is stored into an index routing data structure. Each index routing data structure holds an information about the near and far interest peers along with IP addresses. At rewiring protocol, the peer sends a packet message with an interest centroid randomly to the neighbours for finding similar interest. The highly related neighbours in content interest response with similarity values and IP addresses. Through sending a query to the peers with high similar interest to it, iCluster approach enhances the retrieval effectiveness (Recall) and network traffic costs (70% compared to the flooding approach). The dynamic behaviour of P2P network and looking up information for interest peers are the main challenges of the approach.

- SOON (Li and Vuong, 2008) is a scalable, self-organised overlay network constructed based on ontological semantic clusters of the peers in a small-

world topology. A peer shares an ontological and refined summary of its content as a semantic meta-data information. A peer joins and connects to other peers by the semantic similarity of its distinct semantic concepts with the other peers' semantic concepts. The experimental results show an effective retrieval quality and reduction in information discovery costs using SOON network topology.

- Zhang (2011) proposed a trust-aware P2P-IR system to retrieve -based on two document selection criteria- relevant and trustworthy documents, for a given query as a relevance for user's information need and security perspective. The system was evaluated on two cooperative and uncooperative structured P2P-IR scenarios. In particular, a set of trust factors is identified in the context of P2P-IR, where a content trust model is proposed to estimate the document and document provider trust values for a specific query. The content trust models along with relevance-based ranking both for document ranking and peer selection are combined into a relative weight between relevance and trustworthiness. The combined model maximises the number of relevant documents and at the same time minimises the risk of reviewing untrustworthy documents. The experimental results show the robustness of the proposed model to reduce significantly the possibility of untrustworthy documents at the top-ranked result list.

- Rudomilov and Jelínek (2012) proposed a class-based query routing approach in decentralised unstructured P2P networks. In the network topology, a new peer groups its local documents into semantic class vectors as the vector space model of the document and term weights (term frequency) using online spherical k-means algorithm. This peer then uses Gnutella bootstrap mechanism to send a random walk query message periodically to join the similar semantic peers. The similarity is computed using cosine similarity on the class vector attached to the random walk query and the class vectors of requested peers. The class vectors of all peers are disseminated across the network as a global statistic information for searching processes. To find relevant documents, the requesting peer floods its query to local connected semantic peers and randomly send it to the long distance

peers.

- BloomCast (Chen et al., 2012) is a novel replication strategy to support efficient and effective full-text retrieval in hybrid unstructured P2P networks. The idea behind the Bloomcast is to replicate a set of documents randomly across optimal peers to achieve guaranteed recall. However, the architecture of the hybrid unstructured P2P network consists of three types of peers: normal peers, structured (DHT) peers, and bootstrap peers. The bootstrap peers organise the topology of the network and identify the willingness peers to be DHT peers and the other new peers as normal peers. The DHT peers organise their normal peers and stores information about their states for random peer sampling and network size estimation. Hence, each normal peer generates a compressed form of its documents through inserting sequentially their terms into a bloom filter data structure using a set of hash functions along with their URLs.

  In order to replicate documents, the normal peers need to know the network size to estimate the number of replicas; which is mathematically estimated as a square-root of the network size, and sample a set of random peers to deploy the bloom filter of the documents. Through deploying a lightweight DHT data structure, the normal peers estimate the network size by aggregating the density of peers in ID sections of the DHT peers, which is also used to acquire a set of random peers. In searching process, the encoded language form of runtime query is evaluated against the bloom filters of the replicated documents where the attached URLs of the matched bloom filters contain all the query terms returned as query results. The BloomCast approach was evaluated on a simulated hybrid unstructured P2P network built on a well-known collection called TREC WT10g of around 1.6 million documents. The results show significant query recall of 91% and search latency to 57%. This approach suffers from the problem of aggregating the global information and the query matching mechanism that based on the keywords matching.

- Ke and Mostafa (2013) studied the importance of directing toward better and scalable alternative to IR architecture. The experiments run on decen-

tralised IR operations on various scales of information networks and focus on the impact of network structure on search performance and scalability limit in large information networks. On several experimental settings such as network size, cluster exponent and search methods, a consistent phenomenon is proposed which is called clustering Paradox, in which the level of network clustering imposes a scalability limit.

3.4.5 Reputation-based Systems in Peer-to-Peer Networks

The nature of P2P networks is to give the users full control over what they want to share without restrictions imposed on the quality of published information and behaviours. Hence security issues have been considered to tackle adversary users behaviours over the P2P networks and also preventing them from providing bad quality information (Jøsang et al., 2007). Trust management systems have been emerged to achieve such security issues from acquiring evidence (or trusts) on quality behaviour of users. They have been attention in the development of modern open decentralised P2P systems and have their own root in authentication and authorisation (Kamvar et al., 2003; Xiong and Liu, 2004; Wang and Vassileva, 2004; Wang et al., 2006; Chen et al., 2007; Gómez Mármol et al., 2009).

Another concept that is related to trust is called reputation. The reputation concept is known as the opinions of others on an entity using past direct or indirect interactions. Reputation exists in on-line communities and derived from the underlying social network (Jøsang et al., 2007) as an indication of trustworthiness if an individual entity might have less information to determine the trustworthiness of others. In P2P networks, prior research focuses on the trustworthiness (or reputation) of peers to interact with from security perspective through punishing the malicious peers and preventing the untrustworthy peers from providing untrustworthy documents in the network.

Information retrieval systems in search engines or even in P2P overlay networks seek to retrieve a set of relevant documents for a given query, but most of these documents might be untrusted or duplicated which dilutes the focus of search and leads to trouble the user to precisely and speedily finding the relevant documents (Robert and Sendhilkumar, 2013). However, reputable and relevant documents have been witnessed an important issues in information retrieval in

both search engines and P2P networks. As the motivation of retrieving high-quality and reliable documents, the concept of trust must be formalised by information retrieval systems. Here, I explain a set of related work on trust documents formalization.

Kazai and Milic-Frayling (2008) proposed a reputation-based trust management system to evaluate items in social information retrieval environments (Goh et al., 2007). They proposed a function to calculate trust value for an item derived from the approved votes and the reputation of voters. The voters can be reliable human, reviewers, and a citation network. Their proposed work is highly abstract and there is no experimental evaluation.

Huang et al. (2010) proposed a novel social model for finding a friend of common interests in Online Social Networks (OSN) using trust and popularity values. The trust value is calculated by an algorithm estimating the shortest path between two nodes using trust threshold value, and the popularity value is estimated by using page rank algorithm (Page et al., 1999). The final search output is formulated as a combination of these two values. Their model did not take into account the content trust value, but can be used to determine the trustworthiness of a document provider to retrieve trustworthy documents.

Robert and Sendhilkumar (2013) proposed a provenance model that based on provenance information of an item, which is information of the item's past history, to determine the quality, reliability and amount of trust as factors for identifying trust search results. Their model based on six factors of provenance information which are *who (has authored a document)*, *what (is the content of the document)*, *when (it has been made available)*, *where (it has been published)*, *why (the purpose of the document)*, *How (it is lined)* to ensure that the most trustworthy result at the top of the search results as well as to reduce the untrustworthy results. They construct a provenance matrix that encompasses the six factors and perform an inference over the matrix to calculate the result item score.

Pattanaphanchai et al. (2013) proposed objective trustworthiness criteria to assess the credibility or quality of web information. The trustworthiness criteria extracted from prior studies as a practical criteria that can be adopted to implement an assistance tool to support users' judgements in evaluating the trustworthiness of web information. The proposed normative trustworthiness criteria

are, *authority* relates to the author's identification and credentials, *accuracy* relates to error-free expression of information, *currency* relates to how up-to-date the web information is, and *relevance* that relates to how well the content meets the user's needs. In addition, these criteria were validated by elicited the opinions of an expert panel using a questionnaire and then analysing their responses. The questionnaire is built based on thirteen components from the trustworthiness criteria to allow an expert to rate the effect of the component items on the evaluation of trustworthiness of web information. As a result, tenth of component items were selected as useful components for evaluation. Their proposed work is highly abstract and there is no experimental evaluation.

The previous approaches focused on the security aspect and used the trust (or reputation) concept to protect the users from malicious peers and bad quality documents. They did not exploit these concepts for improving the retrieval effectiveness (or retrieval quality). Hence, in this thesis, I use the reputation concept to enhance the query routing performance in which subsequently the retrieval quality is improved. I believe that this is the first work that focus on using the reputation concept to enhance the effectiveness of retrieval quality by improving the query routing process in P2P-IR systems. In particular, I leave the security aspect for future work and discuss how to exploit the reputation as a metric value for documents and then for the peers to route a query to relevant peers that most likely to contain relevant documents in semi-structured P2P-IR networks.

3.4.6 Discussions

Federated search systems are seeking to mitigate the distributed information retrieval challenges to improve the retrieval quality. A plethora of prior research focused on these challenges and have attempted to provide solutions and suggestions for the sake of enhancing the retrieval effectiveness and efficiency in such systems. In spite of these useful research, the retrieval performance still need to be improved and encourage the researchers to spend more efforts to achieve a high quality P2P-IR systems. In this chapter, however, I discussed the federated search systems especially P2P-IR networks to draw the motivation in using them for information retrieval. As discussed before, P2P-IR networks can be seen as alternative frameworks due to their ability in solving the ethical and technical

issues of centralised search engines. In addition, P2P-IR architectures have a set of advantages clarify the importance of selecting them as IR form, which include (Zhang, 2011; Tigelaar et al., 2012): (i) P2P networks provide more opportunities for using reputation and other selection criteria as well as minimise, especially in decentralised P2P networks, the chance of having a single point of failure, (ii) P2P networks avoid bottlenecks, such as traffic overload, because they can distribute data and balance requests across the network without using a central server, (iii) P2P networks scale-up in processing power and storages due to unused resources of each peer and joined new peers, (iv) Effectiveness and efficiency in updating information at each local peer, and (v) Low maintenance cost of building expensive infrastructures. The P2P-IR networks in searching relevant information suffer from a set of deficiencies, which include: (i) Document retrieval is not efficient due to the different distribution of documents over peers, (ii) Network traffic complexity, (iii) No guarantee of the reliability and quality of documents, and (iv) Comprehensive search that is restricted to a specific part of the network. Although the shortages in using P2P-IR networks as information retrieval framework, I believe considering their solutions to the ethical and technical issues of centralised systems and the mentioned advantages along with the full control of sharing information by users, P2P network architectures are a preferable choice and solution for information retrieval settings.

3.5 Chapter Summary

In order to follow the objectives of improving the web search, this chapter presented prior related research and explained the motivation of using distributed systems for Information Retrieval. The concept of federated search (or distributed information retrieval) was discussed as an alternative solution to tackle the ethical and technical drawbacks of search engines. I also discussed two environments of federated search that are categorised based on the cooperation of resources in providing their statistical information for resource selection process. This chapter focuses on the cooperative resource selection methods as they are the main approaches through out this thesis and one of the challenges in federated search systems. In particular, I discussed resource selection methods that depend on the

statistical information which are gathered from the resources as an indication of relevancy in addition to the state-of-the-art classification-based resource selection approaches. The classification-based resource selection approaches build a classifier model to score and rank resources using training set of features and labels that are related to a specific query and resource. In this thesis, the classification-based resource selection methods are used on semi-structured P2P-IR networks and meta-search environments to validate their effectiveness to be used as baseline methods. I believe some of cooperative resource selection methods are still not used for P2P-IR networks. Finally, due to the topic of this thesis, I thoroughly clarified P2P networks and surveyed some important research work in using Information Retrieval (IR) application under such networks. In particular, the concept, architectures, topologies, and IR application of P2P networks were discussed in more detail. I also discussed the reputation-based systems in P2P-IR networks and how they were used in security aspect to protect the network through preventing the malicious peers from harming the system. In prior research, the reputation concept was used for security perspective by aggregating the rating of users (or opinions) on an entity such as user, item, service, etc., and using them to formulate the reputation or trustworthiness of an entity. However, in this thesis, I will use the reputation concept and formulate it as opinions feedback to improve the query routing in semi-structured P2P-IR networks.

Part II

Clustered Peer to Peer Information Retrieval and Cooperative Query Routing Methods

Semi-structured P2P-IR networks obtain intrinsic characteristics in grouping peers into similar content-based interest and efficiently routing a query to the most likely relevant peers. In Chapter 4 of this part, I discuss the process of building the network architecture and analyse different important design considerations. In Chapter 5, I use the most effective topology in Chapter 4 and study the effectiveness and efficiency of different proposed and suggested query routing and resource selection methods.

Chapter 4

Semi-structured Peer to Peer Information Retrieval

"The brain is wider than the sky."
— Emily Dickinson, (1830-1886)

4.1 Introduction

An effective, navigational topology of P2P network is to combine the peers with similar domain interest into the same cluster (van Rijsbergen, 1979; Doulkeridis et al., 2010; Hai and Guo, 2010). This is because, the peers can reach each other via shortest paths as small-world networks (Watts and Strogatz, 1998; Kleinberg, 2000). In such scenarios, it is worth using a clustering approach to form a small-world network to increase the search efficiency and to reduce the message routing overhead (Klampanos and Jose, 2003). Semi-structured P2P overlay networks are a cluster-based topology which exploits the heterogeneity of nodes with regards to their robustness and capacity to fairly distribute labour on the system. This network is proposed as a promising structure to build retrieval approaches, which contains two types of peers; super (or hub) and regular peers (Klampanos and Jose, 2007). The super-peers have a high level of willingness to store the meta-data of their associated regular peers and communicate with each other to cast queries on behalf of their own regular peers (Androutsellis-Theotokis and Spinellis, 2004). Hence semi-structured P2P networks combine the advantages of the two centralised and decentralised P2P overlay networks in load balancing between the super and regular peers and through providing heterogeneity across

peers to improve the performance (Yang and Garcia-Molina, 2003).

In this chapter, I build a set of semi-structured P2P architectural models with different settings to study the performance boundaries and the effectiveness of their retrieval quality. The motivation of the study is to evaluate different design considerations on retrieval effectiveness to see how far the system is from the centralised model. In particular, I study the performance of three cluster-based semi-structured P2P-IR models and illustrate the effectiveness of several important design considerations and parameters on the retrieval performance, as well as the robustness of these types of networks. The parameter settings include (i) three cluster-based topologies, (ii) different number of super-peers, (iii) failure and departure of regular peers in the system, and (iv) some information retrieval models. The study could be summarised based on the following research questions:

- I propose to use three different clustering approaches for content organisation and compare their performance.
  RQ-4.1: How do different clustering approaches compare in terms of performance for the purposes of semi-structured P2P information retrieval?

- In centralised IR systems, we have a clear understanding of how the state-of-the-art retrieval models behave. However, I conjecture that due to variations in the number of documents within peer's collection, the retrieval effectiveness will vary with centralised systems.
  RQ-4.2: How does the retrieval effectiveness vary with respect to various retrieval models in P2P testing?

- Though the semi-structured P2P systems are a compromise between centralised and decentralised P2P systems, nobody has studied the effect of varying the number of super-peers in the system.
  RQ-4.3: How does changing the number of super-peers in semi-structured P2P-IR systems affect the retrieval performance?

- Tigelaar et al. (2012) provide an overview of the challenges in P2P retrieval systems and present the need to study the churn rate; that is, the effect of leaving peers on retrieval effectiveness.

RQ-4.4: How robust are semi-structured systems to failure or peer departure on retrieval effectiveness?

The remainder of this chapter is organised as follows: Section 4.2 discusses the dataset used to evaluate the proposed topologies and the parameter settings. Section 4.3 creates three semi-structured P2P-IR topologies and compares them with a centralised system of the same collection. Section 4.4 studies the effectiveness of some retrieval models in the proposed semi-structured P2P-IR system. Section 4.5 shows the effectiveness of using different super-peers on semi-structured P2P retrieval networks. In Section 4.6, the robustness of semi-structured P2P-IR systems is studied, followed by the conclusions in Section 4.7.

4.2 Dataset Overview

This section discusses document representation, testbeds and evaluation metrics used to evaluate the topologies along with some parameter settings on semi-structured P2P-IR network; where the documents are the core of the system to be retrieved, the testbeds clarify the environments and their peers, and the evaluation metrics determine the effectiveness of the systems in terms of document retrieval. In addition, the experimental settings and parameters used in the evaluation will be clarified such as the retrieval model for all peers and the merging algorithm used for fusing results.

4.2.1 Document Representation

Documents in the experiments are represented as a vector of terms along with their Term Frequency-Inverted Document Frequency ($TF \cdot IDF$) weight that is defined in Equation 2.2 on page 24. $TF_{t,d}$ is the number of occurrences of term t in document d, the $IDF_{t,c}$ measure for a term t, that is also explained in Equation 2.1 (page 23), is computed as follows:

$$IDF_{t,c} = \log(\frac{N}{DF_{t,c}}) \tag{4.1}$$

where N is the total number of documents in a collection, $DF_{t,c}$ is the number of documents in the collection c that contain a term t. $TF \cdot IDF$ combines the definitions of term frequency and inverse document frequency to produce a

composite weight for each term in each document as:

$$TF \cdot IDF = TF_{t,d} \times IDF_{t,c} \tag{4.2}$$

The $TF \cdot IDF$ weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus (Salton et al., 1975). The power of documents depend on the high frequency of the word in these documents to relatively its decreasing frequency in the corpus.

4.2.2 Testbeds and Testbeds contents

The evaluation depends on large-scale testbeds suggested as a real baseline to evaluate the P2P-IR models (Klampanos et al., 2005). These testbeds are developed based on TREC WT10g collection(11680 web domains)[1] as one of the public evaluation corpus for Information Retrieval (IR). The TREC WT10g is a 10 gigabytes corpus, which contains 1,692,096 English web documents used for the evaluation of Information Retrieval systems. The P2P-IR testbeds have three properties, which are: (i) each peer shares a limited number of topics; (ii) the documents are distributed in a power-law pattern; (iii) and availability of content replication across peers in specific testbeds.

The individual testbeds are designed to address a number of P2P-IR environments through different document distributions and concentrations of relevant documents. The testbeds can be categorised into three different environments, which are information sharing environments ($ASIS^*$ family), Digital library environment (DL^* family), and uniformly distributed environments (U^* family) (Klampanos et al., 2005). Each environment contains two testbeds of peers' collections based on replicating documents across peers, which are referred as WOR (**W**ith**O**ut **R**eplication) and WR (**W**ith **R**eplication). The ASISWOR (**ASIS** **W**ith**O**ut **R**eplication) collection represents each peer's collection by one domain. The DLWOR (**D**igital **L**ibrary **W**ith**O**ut **R**eplication) selects the first 1,500 largest domains as attractors for other domains by using the cosine similarity measure between homepages of the related domains. The UWOR (**U** **W**ith**O**ut **R**eplication) divides the available web domains into three buckets; which are undersized, oversized and properly sized buckets according to the num-

[1]Text REtrieval Conference (TREC) WT10g (10 gigabytes): http://ir.dcs.gla.ac.uk/test_collections/wt10g.html (October, 2016)

ber of documents they share. In particular, each excessive document is moved from the oversized bucket into its closest undersized domain using the cosine similarity between the moved document vector and the homepage of each of the undersized domains. Once the undersized bucket or the oversized bucket reach the desired number of documents, they are moved into the properly sized bucket. On the other hand, all testbeds with replication; ASISWR, DLWR, and UWR replicate relevant documents by pulling them randomly into a peer's collection through exploiting inter-domain links between web domains.

Table 4.1: Test-beds general properties

Characteristics	ASISWOR	ASISWR	DLWOR	DLWR	UWOR	UWR
# Peers	11,680	11,680	1,500	1,500	11,680	11,680
# Docs	1,692,096	1,788,248	1,692,096	1,740,385	1,692,096	1,788,896
Max.Peer_Docs	26,505	33,874	26,505	33,874	145	7,514
Min.Peer_Docs	5	5	171	174	140	8
Average.Peer_Docs	144.87	153.1	1128.54	1160.26	144.87	153.16
#Peers.Relevant_Docs	12.28%	12.91%	55.93%	56.73%	29.60%	30.36%
Max.Peer_Relevant	61.54%	57.14%	15.14%	15.14%	17.24%	14.48%
Average.Peer_Relevant	0.31%	0.33%	0.41%	0.43%	0.43%	0.35%

Table 4.1 provides some testbeds' statistics in the different three environments. #Peers refers to the number of peers in each testbed; 11,680 peers for $ASIS^*$ and U^* families while 1,500 peers for DL^* family. #Docs represents the number of documents in the testbed as a whole collection, which is the same for testbeds without replication with 1,692,096 docs and varies for the other testbeds with replication. Max.Peer_Docs and Min.Peer_Docs are both for the peers that contain a maximum and minimum number of documents in the testbed respectively, whereas Average.Peer_Docs determines the average number of documents across the peers in a specific testbed. In terms of relevant document distribution, #Peers.Relevant_Docs represents the percentage of peers that contain relevant documents in the testbed, which is higher in DL^* family than $ASIS^*$ and U^* families as it has a small number of peers. The U^* family contains more peers with relevant documents than $ASIS^*$ family because of uniform distribution of the documents across peers, which makes more relevant documents to be disseminated over several peers. Max.Peer_Relevant refers a maximum percentage of relevant documents of a maximum relevant peer with respect to the whole doc-

uments. Average.Peer_Relevant represents the average percentage of relevant documents of peers to all documents in testbed as a whole collection.

4.2.3 Query Set and Relevant Judgements

The standard query set for the TREC WT10g corpus is TREC topics 451-550[1], which is provided by the US National Institute for Standards and Technology (NIST). Basically, TREC topics consist of three fields of <title>, <description> and <narrative> where Table 4.2 shows three examples. According to the study of users' query behaviour in the web, it has been observed that the average query length for text retrieval in web search engines is less than 2.23 words (Nguyen et al., 2007; Lewandowski, 2015) with 67% over all posted queries (Jansen et al., 2000a). Therefore, the <title> field in each TREC topics 451-550 is selected as the query set in the experiments because the average length of <title> field is 2, which is close to the average query length in real P2P networks.

Table 4.2: Examples on TREC topics 451-550

Topic	460	461	539
<title>	Who was Moses?	lava lamps	authors who suffered from depression
<description>	Find documents that discuss the biblical figure of Moses.	Find documents that discuss the origin or operation of lava lamps.	Which authors suffered from depression?
<narrative>	A relevant document includes any information concerning Moses and his deeds regarding the Israelites.	A relevant document must contain information on the origin or the operation of the lava lamp.	A relevant document will name authors who were depressed.

4.2.4 Evaluation Metrics

The performance of the models is measured by retrieval accuracy. Retrieval accuracy will be measured according to the Precision, Recall, P@10, P@30, P@100, Average Precision and Mean Average Precision (MAP) metrics, which are clarified in Chapter 2. In IR systems, the users assume that the relevant documents are at the top of result list, therefore they interest with the top-ranked documents

[1]TREC English Test Questions (Topics):http://trec.nist.gov/data/webmain.html (October, 2016)

rather than the whole results set (Buckley and Voorhees, 2004), I use five measurement values that depend on retrieved relevant documents at a cut-off value; P@10,30,100, Average Precision, and MAP. P@n metric measures the fraction of relevant documents at a top cut-off value n of the ranked-based results set, which are usually given by the top-k retrieved documents. As n increases search systems struggle to maintain the corresponding search precision (Kulkarni and Callan, 2015). This will motivate us to include metrics such as P@10, P@30, and P@100 to test the effectiveness of the systems at deeper ranks. A related approach that has been used more frequently in recent times is MAP, where the precision is measured at every point at which a relevant document is obtained and then averaged over all relevant documents to obtain the Average Precision for a given query. For a set of queries, the mean of the average precision for all queries is the MAP of that IR system.

4.2.5 Experimental Settings and Parameters

In order to evaluate the semi-structured P2P-IR network, I did a set of experiments along with specific unified parameter settings for all participating peers. Each peer in the systems has a retrieval model which is used for extracting relevant documents based on a user query. I used a well-known information retrieval platform, which is called TERRIER (Terabyte Retriever) as an indexer and a search interface for the whole peers (Ounis et al., 2006). In the experiments, I used Okapi (BM25) retrieval model (Jones et al., 2000), which is a probabilistic retrieval function that ranks a set of documents based on the query terms in each document. The Okapi model is based on term frequency and document length. The Okapi BM25 basic weighting is calculated as follows, which is also defined in Equation 2.16: Given a query Q, containing terms $t_1, ..., t_n$, the BM25 score of a document d is:

$$\text{score}(d, Q) = \sum_{i=1}^{n} IDF_{t_i,c} \cdot \frac{TF_{t_i,d} \cdot (k1 + 1)}{TF_{t_i,d} + k1 \cdot (1 - b + b \cdot \frac{dl_d}{\text{avgdl}})}, \tag{4.3}$$

Where $TF_{t_i,d}$ is t_i's term frequency in the document d, dl_d is the length of the document d in words, and $avgdl$ is the average document length in the text collection from which documents are drawn. $k1$ and b are free parameters, in the experiments the two values are $k1 = 1.2$ and $b = 0.75$, which are the default

values for the model.

$IDF_{t_i,c}$ is the IDF (inverse document frequency) weight of the query term t_i. It is usually computed as:

$$IDF_{t_i,c} = \log \frac{N - df_{t_i,c} + 0.5}{df_{t_i,c} + 0.5}, \tag{4.4}$$

Where N is the total number of documents in the collection, and $df_{t_i,c}$ is the number of documents containing t_i in corpus c.

I fixed all the numbers of retrieved results per super-peers and their regular peers with 1,000 documents per result. In the experiments, I used a well-known merging algorithm called CombMNZ algorithm (Shaw and Fox, 1994). The CombMNZ is an unsupervised merging algorithm which is simple, effective and well-studied (Lee, 1997). In particular, the CombMNZ uses the CombSUM method as one of its parameters to sum all scores of a document and then multiply this value by a number of the non-zero document scores to reward the documents that have a high score and appeared in multiple lists as shown in Equation 4.5 and 4.6. The assumption behind the CombMNZ algorithm is that a document retrieved in more than one result is better than another document that has the same similarity or rank order retrieved in a single result (Lee, 1997). The major advantages of CombMNZ method are: (i) it does not require document processing, it only requires rank score normalization, (ii) it does not require a similarity score, which is not available in most popular search engines, it requires only retrieved documents with rank order, (iii) it is simple and requires little processing time and disk space. I used CombMNZ in the experiments in two levels as the regular peers' results are merged by their super-peer and the super-peers' results are merged at other super-peers.

$$CombSUM(d) = \sum_{i=1}^{n} s_i(d) \tag{4.5}$$

where $s_i(d)$ is the score that the peer i assigns to document d. If peer i does not have the document d the value of $s_i(d) = 0$.

$$CombMNZ(d) = m * CombSUM(d) \tag{4.6}$$

where m refers to the number of result lists in which document d appears.

4.3 Semi-structured Cluster-based P2P-IR Network

There are a variety of architectures that vary in terms of the organisation of the peers. In semi-structured P2P retrieval systems, two layers of clustering are used for organising the content in a coherent way.

4.3.1 Semi-structured Cluster-based P2P-IR Architecture

The cluster-based architecture uses two levels of clustering, as specified in (Klampanos and Jose, 2007, 2003). Figure 4.1 illustrates as an example the construction of the clustered 2-tier architecture. Each of the peers maintains a subset of documents, as shown by the different P_is on the left side of the figure. The subset of documents within each peer is subjected to a clustering process, illustrated in the figure as Step **A**; I will call this *intra-peer clustering*. Though the figure shows 3 clusters consistently for every peer, there could, in general, be any number of clusters. Phase **B** clusters these intra-peer clusters, across peers, into a specified number (two, in the figure) of clusters. Each such cluster is managed by a super-peer (SP_i). Due to the clustering, not every super-peer necessarily would have representation from each peer; in the example, SP_2 does not have representation from P_1. The super-peer level, as may be noted, is an additional layer, giving the framework the name 2-tier. Every query to the P2P-IR system is sent to *each* of the super-peers, which would then use the information from the intra-peer clusters it manages, to route the query to one or more peers to which it is connected.

For a news search engine where different peers manage content from separate news agencies, sports related news articles may be separated out from others within each peer in the intra-peer clustering phase. Due to the second-level clustering, the sports clusters from separate peers are expected to be collected into a super-peer. Thus, the 2-stage clustering process ensures that routing decisions can be made at the level of super-peers that manage coherent content internally as well as among each other, while not disturbing the document assignment to peers; this likens the scenario to a domain-specific search engine at each super-peer. This section, in particular, discusses the two clustering steps used in this

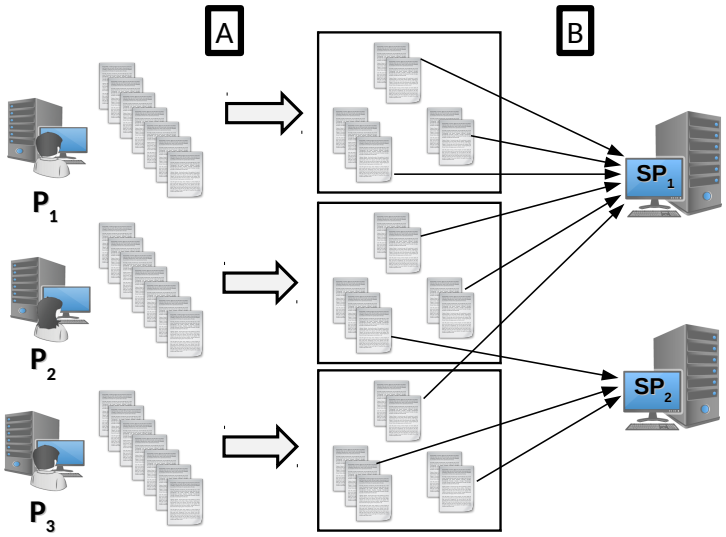


Figure 4.1: Clustered 2-Tier Architecture

thesis to build the semi-structured P2P-IR network as an effective, navigational topology for a distributed information retrieval system.

4.3.1.1 Intra-peer Clustering

In the P2P networks, the peer holds a limited set of topics which motivates to group the peers' documents that have the same content topics to be used as a peer representation and to build the super-peer network (Klampanos and Jose, 2007). In this phase, the documents within each peer are clustered to form lexically coherent topical clusters. I use bisecting K-means clustering algorithm to perform intra-peer clustering on TF-IDF document vectors. The bisecting K-means algorithm splits a cluster into exactly two clusters; this is repeated in parallel as many times as needed within each peer until the resultant clusters no longer satisfy a maximum cardinality constraint. Particularly, the number of iterations for all parallel K-means clustering algorithm is 40 iterations and the final clusters in leaf node level depends on the stopping conditions used in the algorithm. The stopping criteria for the bisecting K-means occur when one of the splitting clustering algorithm (parallel K-means) can not discover any new cluster

(or each cluster has 5 documents) or the number of documents in the new cluster is 20% of all documents that used to construct such a cluster. The stopping criteria is used due to the small number of documents in each peer in order to build a coherent clusters. I use the term *peer clusters* to refer to the resultant clusters. Each peer cluster may be represented in the vector space model using the mean of the vectors associated with the documents in the cluster as peer clusters' centroids.

It may be noted that any clustering algorithm can be used to identify peer clusters; the choice of bisecting K-means is driven by efficiency considerations (Steinbach et al., 2000). Bisecting K-means clustering algorithm is considered as one of the divisive (or partition) hierarchical techniques that start with a single cluster of all the documents. Bisecting K-means clustering algorithm exhibits better clustering performance in comparison to the standard k-means approach, and similar performance to hierarchical approaches (Steinbach et al., 2000). In addition, bisecting K-means has a lower computational complexity than agglomerative hierarchical clustering techniques- O(n) versus O(n^2). The cluster-based architecture, unlike other architectures, has an elaborate set-up phase that involves clustering of documents. Thus, an understanding of the computational and memory costs of the clustering phase is important in analysing the applicability of routing algorithms that work upon it.

Table 4.3: The Bisecting K-means Clustering analysis

Meta-Info	No.of.Clusters per Peer			Av.Docs.per Peers' Clusters			Av.Terms.per Peers' Clusters			Clustering Time (in secs)		
Test-beds	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
DLWOR	1	119	12.3	304.2	601.1	371.3	4,090.4	9,172.5	5,360	2.26	349.9	14.9
DLWR	1	137	12	15.2	268.7	628.7	359.3	163.4	4,120.4	9,403.5	5,497.8	2,214
ASISWOR	1	166	3.9	39.2	68.5	46.8	853	1,713.5	1,129.3	0.1	469.1	2.6
ASISWR	1	154	4	6.7	40.5	71.4	48.1	13.4	915.6	1,840.3	1,202.2	405.4
UWOR	1	23	6.2	4.6	26.5	54.8	35.5	13.9	1,293.2	3,126	1,891.6	827.2
UWR	1	29	6.2	4.7	30	61.3	40.1	15.8	1,400.2	3,332.1	2,037.3	884

The statistics of the bisecting K-means clustering algorithm appear in Table 4.3. The table details the number of clusters, documents per cluster, the number of terms per cluster, and the time taken for clustering on average. As shown, the values are different in file sharing ($ASIS^*$) from digital libraries (DL^*) and uniformly distributed environments (U^*). Digital library environment has a large

number of centroids with higher values on average of peers' centroids. As the small number of peers in this environment have a large number of documents with 1,128 and 1,160 on average for DLWOR and DLWR respectively. The other two environments show opposite values of digital library environment with a small number of centroids on average peers' centroids. In addition, the clustering is seen to just take a few seconds to complete on average, and a few minutes at worst; I presume that these would be regarded as very low overheads for the setup phase. The bisecting K-means clustering algorithm was seen to be even faster, since the clustering is done on the peer clusters that were seen to be fewer than the number of documents per peer on average.

4.3.1.2 Inter-peer Clustering

In this phase, I propose to use three other alternative clustering approaches to constructing the super-peer level of the networks, which are: (i) K-means, (ii) Half K-means Single-pass, (iii) and Approximation single-pass architectures. The centroids of all peers that resulted from bisecting K-means in the first level were used as input for three scenarios to construct the super-peer networks in different ways.

K-means Architecture: The K-means clustering algorithm extracted 50 centroids (k is determined manually to 50) as output vectors to be the super-peers' content descriptions. The super-peer content descriptions in the systems are used for routing query and to be a descriptive information of all their regular peers. K-means architecture has a benefit of building the network by selecting the better centroids at each iteration and combines the arrived centroids around the constructed super-peers of the same topics.

Half K-means single-pass Architecture: The second scenario runs K-means clustering algorithm on half of the peers' centroids with the same settings of the previous scenario to build 50 super-peers and then applies single-pass clustering algorithm on the other half of the centroids. The cosine similarity was used as a measurement between centroid vectors on a threshold of 0.5 as the cosine similarity of TF-IDF vectors is between the interval [0,1]. The half K-means single-pass algorithm assumes that half of the resources in the network are available and

they can be used as attractors for the other arrived peers to discover the related semantic groupings for joining the system.

Approximation single-pass Architecture: As the single-pass clustering is computationally expensive, I use an approximation single-pass approach, which is executed on a distributed Hadoop cluster of 8 nodes. In the approximation single-pass method, I divided the peers into eight packets and then used the single-pass algorithm on each packet to create super-peers for each packet. The super-peers in all packets were used as a topology from the assumption that the super-peers might be created separately as independent components from each other.

The architectures of the three scenarios follow a small-world network with network diameter (or longest shortest path) of three between network nodes. The super-peers are connected with a set of peers and contains a set of topics. Each super-peer has a centroid which is a description and an indication for the topics of connected peers' contents. In summary, in all architectures, the objective is to organise the contents of the network into high-level categories, which can be used for routing queries appropriately. The resultant k clusters' centroids are handled by a separate *super-peer*. Thus, each super-peer gets a subset of peer clusters' centroids from across peers; since I do not use any constraints in the inter-peer clustering step, each super-peer may receive zero or many peer clusters from a given peer. Hence each peer might be connected to multiple super-peers with different peer centroids, as for example P_3 and P_5 as shown in Figure 4.2. The number of super-peers in a real P2P network depends on the willingness of peers to act as a super-peer.

Due to the two-level clustering involved, the *super-peers* get to manage coherent content from across the peers, and are hence said to be nodal points of *content-aware groups* (Klampanos and Jose, 2007). Each super-peer manages information about multiple clusters. I will denote the k^{th} cluster from peer P_i as P_i^k. For every super-peer SP_j, C_{SP_j} denotes the set of clusters that are managed by SP_j. Due to the clustering-based construction, C_{SP_j} could potentially contain multiple clusters from a specific peer. I use a centroid-based representation throughout; thus, $Cd(P_i^k)$ denotes the centroid of the documents within the k^{th}

cluster in the i^{th} peer. The entry for each word takes the average of the value of the word across the vectors that form the centroids.

$$Cd(P_i^k)[w] = \frac{\sum_{d \in P_i^k} tf.idf(w,d)}{\#docs\ in\ P_i^k} \tag{4.7}$$

$$Cd(C_{SP_j})[w] = \frac{\sum_{P_x^y \in C_{SP_j}} Cd(P_x^y)[w]}{\#clusters\ in\ C_{SP_j}} \tag{4.8}$$

where $tf.idf(w,d)$ denotes the TF-IDF score of the word w in document d.

The centroid-based representation is used in this approach in order to route the given query to the relevant super-peers. Hence the approach determines the relevant super-peers by calculating the cosine similarity between the super-peer centroid representation and the query vector.

In semi-structured P2P-IR networks, the query routing process can occur in two different forms; flooding and resource selection methods. Figure 4.2 shows query routing on the semi-structured P2P-IR network as an example of seven users (peers) and four super-peers. The user of a peer (i.e, sender) initiates and sends a query to the connected super-peer (i.e, SP_1). The super-peer then routes the query to its local peers and other selected super-peers. The query routing techniques select the most relevant super-peers based on the cosine similarity between the super-peers' centroids and the query vector. At each super-peer level, the flooding approach sends the query to all local peers as indicated by the dot-dashed arrow (or red arrow), while the resource selection methods send the query to most likely peers that have relevant documents as indicated by the dotted arrow (or blue arrow). Subsequently, the requested peers execute the information retrieval process on their indices and return a list of results to their connected super-peer. The final result list is compiled at the sender's super-peer by merging results from different super-peers.

4.3.2 Retrieval Effectiveness of Semi-structured P2P-IR Architectures

The retrieval effectiveness of three clustered-based semi-structured P2P architectures are evaluated against a centralised search engine, which is considered as a point of reference. Firstly, I compare the results with a centralised system to see

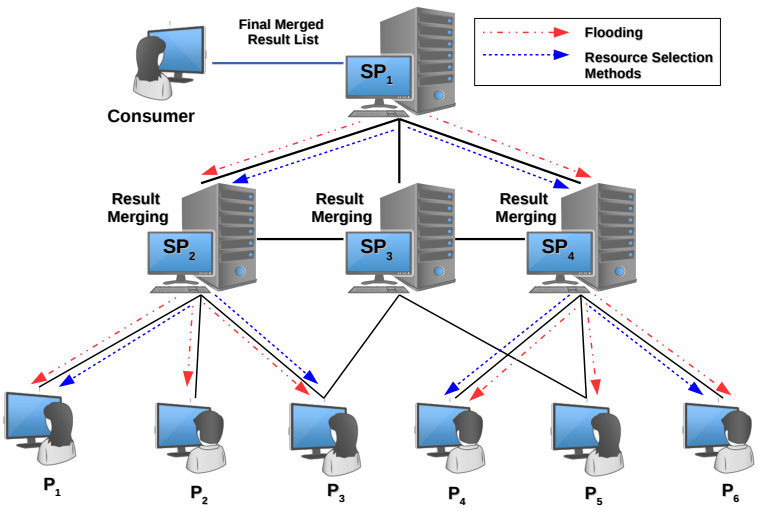


Figure 4.2: Query Routing in Semi-Structured P2P-IR models

which architecture is closer to this reference system, as the results of P2P architectures are not comparable with centralised systems (Xu and Callan, 1998). Hence I do not use the statistical significant test and only exhibit the results of three clustered-based semi-structured P2P architectures on evaluation metrics. Table 4.4 shows the six metrics on IR effectiveness, employing BM25 retrieval model and retrieving 1,000 documents per query, for a centralised system averaged over all 100 TREC topics (451-550) using TREC WT10g collection.

Table 4.4: The effectiveness of retrieval information in a centralised system

Topics	Recall	Precision	P@10	P@30	P@100	MAP
100	0.7008	0.0406	0.299	0.2293	0.1418	0.1903

In the centralised index, the average number of relevant documents retrieved is 38.53 which is 4% of retrieved documents; approximately 3, 7, 14, and 38.53 documents at positions 10, 30, 100, and 203 respectively with 970.5 retrieved documents on average.

4.3.2.1 K-means Architecture

The retrieval effectiveness of K-means architecture can be shown in Table 4.5, which also includes the six metrics of evaluation as in the centralised system averaged over all 100 TREC topics (451-550).

Table 4.5: K-means Architecture effectiveness

Testbeds	Recall	Precision	P@10	P@30	P@100	MAP
ASISWOR	0.2984	0.0161	**0.078**	**0.0527**	0.0317	0.0240
DLWOR	**0.3655**	0.0194	0.0717	0.0502	0.0336	**0.0354**
UWOR	0.3428	**0.0202**	0.0580	0.0510	**0.0381**	0.0243
ASISWR	0.2530	0.0140	0.0280	0.0250	0.0208	0.0138
DLWR	0.2564	0.0134	0.0173	0.0238	0.0198	0.0132
UWR	0.2027	0.0120	0.0150	0.0197	0.0177	0.0104

As shown, there is a substantial difference in retrieval effectiveness between the centralised system and the K-means semi-structured architecture over all testbeds. The reason is back to the document distribution in P2P systems that causes a problem using retrieval models to determine the statistics of terms. On a comparison between testbeds, ASISWOR has the best P@10 and P@30 values, DLWOR testbed obtains better Recall and MAP values, while UWOR testbed gains higher Precision and P@100 values. DLWOR testbed on average gets competitive and better retrieval performance over all testbeds in this architecture. This shows the effect of K-means architecture on the retrieval effectiveness. In general, the distributional nature of P2P systems leads to poor results (approximately 19% versus 3% of MAP values) in comparison with the centralised system. These difference show the effect of replication and the document distribution nature of various real-life applications. In addition, as seen, the testbeds without replication have better retrieval quality more than the testbeds with replication, the reason is that the replication strategy changes the distribution of terms in the collection that has an effect on retrieval and merging processes. This means that if the number of relevant documents increased in a specific ratio in the collection, these documents become non-discriminative. To see how similar this architecture is to the centralised system, I define *a closeness metric that refers to the percentage decrease (or loss) in retrieval quality in comparison to the centralised system*

over all metrics' values. The percentage decrease is also calculated on all testbeds for a specific metric. The closeness value of such architecture is approximately 23.44%, which is averaged over six values of closeness on each metric as 40.88% (or average Recall percentage decrease over each Recall value of a testbed to the centralised system), 39.04%, 14.94%, 16.17%, 19.01%, and 10.61% for Recall, Precision, P@10, P@30, P@100, and MAP respectively.

4.3.2.2 Half K-means Single-pass Architecture

The Half K-means Single-pass architecture assumes that the semi-structured P2P network is already available for the other arrived peers' centroids. The number of centroids for each testbed to create the super-peers are on average 9,000, 23,360, and 35,040 for DL^*, $ASIS^*$, and U^* families respectively. The number of centroids in DL^* is the lowest as the number of peers is 10% (i.e, 1,500) less than the other environments (i.e, 11,680) with approximately 12 centroids on average. The K-means clustering algorithm creates 50 super-peers (k=50) with their centroids as an average of the arriving centroids. Then, a single-pass clustering algorithm is used to connect the other arrived centroids only to these super-peers. The retrieval effectiveness of half K-means single-pass architecture is shown in Table 4.6.

Table 4.6: Half K-means Single-pass Architecture effectiveness

Testbeds	Recall	Precision	P@10	P@30	P@100	MAP
ASISWOR	**0.1354**	**0.0113**	**0.0602**	**0.0405**	**0.0219**	**0.0216**
DLWOR	0.0466	0.0045	0.0598	0.0292	0.0126	0.0145
UWOR	0.0597	0.0082	0.0286	0.0235	0.0132	0.0044
ASISWR	0.1248	0.0108	0.0296	0.0282	0.0173	0.0115
DLWR	0.0512	0.0077	0.0463	0.027	0.0128	0.0053
UWR	0.0498	0.0086	0.0184	0.0153	0.0096	0.0031

Here in Table 4.6, I am showing the effectiveness values for the Half K-means Single-pass architecture in different testbeds. In addition, I compare how close their values are to the centralised system. The closeness of this architecture to the centralised system is approximately 12.18% of 11.12%, 20.98%, 13.54%, 11.90%, 10.27%, and 5.29% on each metric respectively. The result shows that this architecture is closer to the centralised approach than the k-means architecture.

4.3.2.3 Approximation Single-pass Architecture

As explained before, the centroids' descriptions of super-peers were formed by using single-pass clustering algorithm on 8 packets of peers' centroids to create 50 centroids described as super-peers. The result of retrieval effectiveness can be shown in Table 4.7.

Table 4.7: Approximation Single-pass Architecture effectiveness

Testbeds	Recall	Precision	P@10	P@30	P@100	MAP	P@10_Klampanos
ASISWOR	0.3627	0.0202	0.0293	0.0195	0.0236	0.0197	0.0196
DLWOR	0.2892	0.0151	**0.064**	**0.0437**	0.0243	0.022	0.0063
UWOR	**0.4513**	**0.0255**	0.0172	0.0259	**0.042**	**0.0302**	0.060
ASISWR	0.2397	0.0131	0.009	0.0097	0.0142	0.0107	-
DLWR	0.3294	0.0165	0.012	0.0127	0.0161	0.0131	-
UWR	0.2699	0.014	0.009	0.0123	0.0157	0.0133	-

The approximation single-pass architecture shows better performance than half K-means Single-pass architecture for which highest Recall (i.e, 0.4513) is achieved. The UWOR testbed also has high scores for Precision, P@100, and MAP metrics, where DLWOR obtains high values in P@10 and P@30 in such architecture. This architecture is closer to the centralised system than the half K-means single-pass architecture with closeness value of 21.90% on average of the metrics as 11.12%, 20.98%, 13.54%, 11.90%, 10.27%, and 5.29% respectively.

The results can be more fairly compared to other works which use the same evaluation framework. The authors in (Klampanos and Jose, 2007) use the same testbeds and build the network in two level of clustering; ward clustering at peers documents level and single-pass clustering at super-peer level as our target clustered-based architecture discussed in Chapter 3. In their work, they use the P@10 metric at different thresholds of clustering on four testbeds. The best P@10 values are 0.0196, 0.0063, and 0.060 for ASISWOR, DLWOR, and UWOR respectively. In the three testbeds; ASISWOR, DLWOR, and UWOR, the clustering architectures have better retrieval results compared to their approach based on P@10 metric. This means, in terms of effectiveness, the proposed architectures are better and/or comparable to the original proposal in (Klampanos and Jose, 2007) to P@10 values of the testbeds. But, the approximation single-pass archi-

tecture in the semi-structured P2P-IR system, especially on UWOR testbed, has a lower value less than 0.06, which means that their approach outperforms it in such scenarios. The reason is that they tuned the threshold parameter for the single-pass clustering algorithm to get these results.

In summary, on average the best architecture of three scenarios is the K-means architecture, with MAP value of 0.0201 over all testbeds, which is used as a target network for the rest of the thesis. In contrast, the approximation single-pass approach has MAP value of 0.0182, while half K-means single-pass has MAP value of 0.0102. These experiments address **RQ-4.1** and confirm that clustering approaches can affect P2P scenarios differently. The results clearly show the need for developing a robust content organisation methodology. These results are to compare different semi-structured architectures to propose a set of techniques on the target architecture to increase the performance of semi-structured P2P-IR networks. There are many reasons behind the results of these topologies in comparison to the referenced centralised system, which might include:

1. There are some topics (queries) in both centralised and semi-structured P2P systems have poor retrieval results, where the average precision (AP) are zero or less than 0.01 in the systems. In semi-structured P2P systems, there are several topics that have a larger number of non-relevant than relevant documents in a specific peer. Hence, in the semi-structured P2P topologies, a set of peers might have non-relevant documents that contain the topic (or query) terms and do not have relevant documents, refer to Table 4.1 to see the relevant documents distribution; which means the non-relevant documents have higher scores in a peer's collection and leads to a poor result quality as a whole.

2. Errors in the super-peers' centroids that are constructed from a clustering algorithm, which prevent poor topics from reaching the relevant peers for retrieval.

3. Each semi-structured P2P approach has its own characteristics in creating the topology of the system. The way of building these topologies has an impact on retrieval quality due to the distribution of documents and the shape of the network. Therefore, each approach has a better scenario

from an Information Retrieval perspective; DLWOR in a K-mean topology, ASISWOR in Half K-means a Single-pass topology, and UWOR in an approximation single-pass topology. On average, k-means in DLWOR testbed outperforms all semi-structured topologies.

4.4 Retrieval models in Semi-structured P2P-IR Networks

In this section, a set of retrieval models is used to compare the retrieval effectiveness of the semi-structured P2P-IR networks. Retrieval models have characteristics and assumptions, which differ from one another in retrieving relevant documents. However, the comparison between retrieval models in the centralised system was conducted and studied in the literature. Hence someone might assume that the same level of performance is expected in P2P-IR systems. Given that the content of the peers are dynamically changed, it is important to study the retrieval quality of P2P-IR systems in different retrieval models.

The retrieval models that I studied come from different families implemented in the TERRIER framework (Ounis et al., 2006)[1] and explained in more detail in Chapter 2. Figure 4.3 shows the effectiveness of retrieval models on the centralised testbed and the other P2P testbeds. I used the F-score metric as the average value combining the Precision and Recall values. The retrieval models behave in a different manner between testbeds and the centralised testbed, because of the distribution of the terms and documents in P2P-IR systems. The retrieval models behave based on the terms' statistics in the collection and each of them has a specific intuition and parameters.

As shown in Figure 4.3, the language retrieval models, which are DirichletLM and Hiemstra_LM, occupies the best and worst retrieval models in centralised systems with F-score 0.081 and 0.0689 respectively. Recent study shows the better performance of DirichletLM model on the same collection (i.e, WT10g) compared to recently developed retrieval methods (Cummins, 2016). In contrary, the retrieval models in semi-structured P2P-IR models perform differently, where the best retrieval model for testbeds without replication is the LGD retrieval model

[1]The algorithms are also discussed in this website http://terrier.org/ (October, 2016)

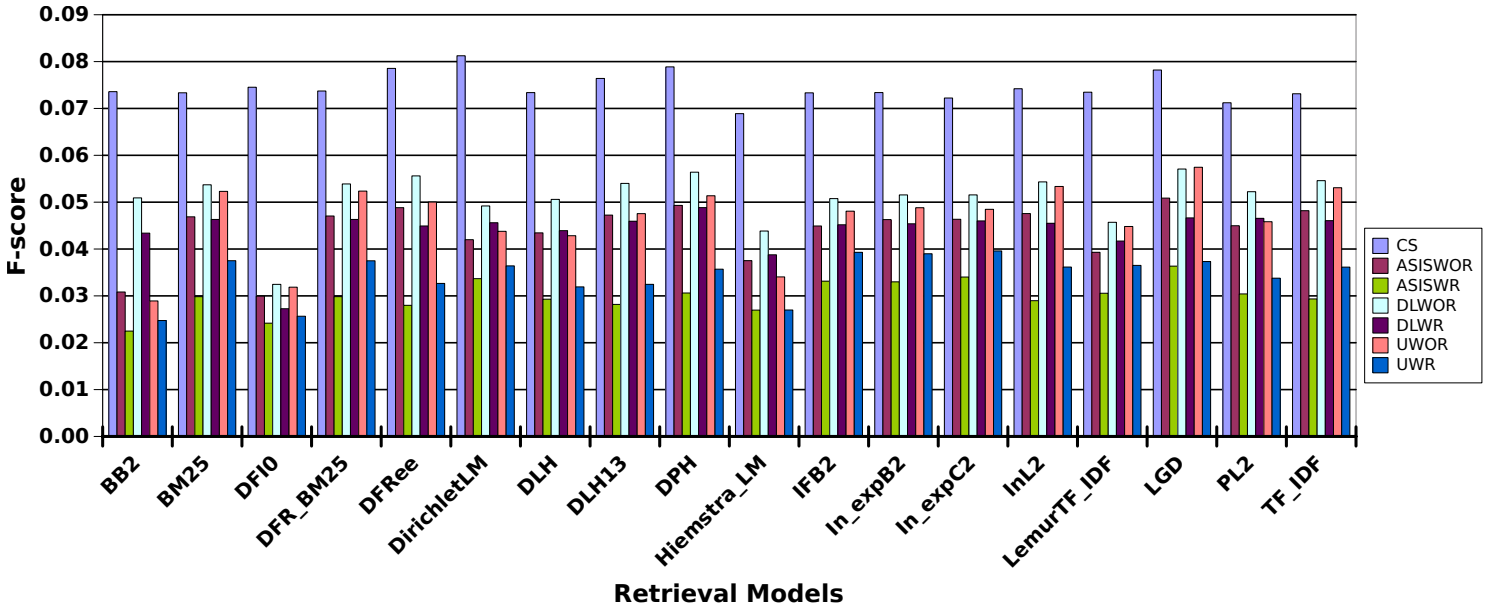


Figure 4.3: Retrieval models over Semi-structured P2P-IR system.

with F-score 0.051, 0.0571, and 0.0574 for ASISWOR, DLWOR, and UWOR respectively. The worst retrieval model on the same testbeds is DFI0 model with F-score 0.030, 0.032, and 0.032 for ASISWOR, DLWOR, and UWOR respectively. The testbeds with replication perform in a different way in comparing with the testbed without replication, because the replicating models replicate relevant documents on different peers and change the distribution of terms for retrieval perspective. The best retrieval models in testbeds with replication are LGD for ASISWR, DPH for DLWR, and In_expC2 for UWR with F-score 0.036, 0.049, and 0.04 respectively, while the bad ones are BB2 for ASISWR, DFI0 for DLWR, and BB2 for UWR with F-score 0.022, 0.03, 0.045 respectively. Ultimately, on average the best retrieval model on all testbeds is the LGD model with approximate F-score value of 0.05, where the worst one is DFI0 model with approximate F-score value of 0.029. I conclude that there are differences in the retrieval effectiveness of retrieval models in the centralised and distributed systems, especially given the heterogeneous distribution of collections in P2P networks; which means that the choice of a retrieval model for a specific peer depends on the collection size of the peer. Since, in P2P networks, the retrieval models' parameters have to be studied carefully in the designing phase (**RQ-4.2**).

4.5 Effect of Different Number of Super-peers

In semi-structured P2P systems, some powerful peers act as super-peer which is exploited for effective and efficient retrieval. However, it is not clear how many super-peers are needed for robust performance. In this section, I study the effect of the number of super-peers on retrieval effectiveness in such networks. Due to the high computational complexity, I experiment with only two testbeds; ASISWOR and DLWOR, that simulate the file-sharing and digital library scenarios. However, I re-implement the second level of clustering by increasing the number of super-peers between 5 and 100. The performance of increasing number of super-peers on two testbeds is shown in Figure 4.4.

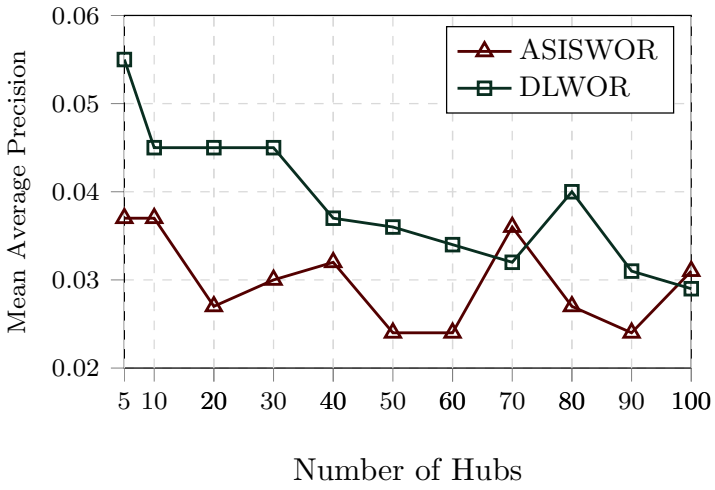


Figure 4.4: The effect of varying the number of super-peers.

Overall, increasing the number of super-peers has an effect on retrieval quality (**RQ-4.3**). As shown in Figure 4.4, I can conclude that the retrieval effectiveness increases as the number of super-peers fall, i.e. when the system tends towards a centralised system. At some point, DLWOR at 80 super-peers and ASISWOR at 70 super-peers, the effectiveness increases which may be an anomaly due to the noise in the clusters. This apparently shows that the centroids' descriptions of the super-peers play a major role in retrieval effectiveness.

4.6 Effect on Robustness

One of the essential challenges in P2P network is the robustness of the system. The dynamic nature of such networks makes the participating peers possess the liberty to share only the content they desire or leave the system at an arbitrary time. Therefore, it is necessary to study the robustness of the system to design a powerful P2P retrieval system to adapt the dynamic behaviour of participating peers, which is referred to *churn rate* (Stutzbach and Rejaie, 2006). However, the robustness can be measured as the percent of available peers in the system at a given time or the percent of available content for retrieval. In this section, I study the robustness of the semi-structured P2P-IR system from the availability of relevant documents in the system and its effect on retrieval effectiveness to meet the users' information need.

Table 4.8: The effectiveness of Flooding method for the Four testbeds (MAP).

DLWOR	DLWR	ASISWOR	ASISWR
0.087	0.0223	0.071	0.0173

The experimental methodology on robustness is as follows: the system selects random numbers of peers, from 10% to 50%, on four testbeds as candidates to be removed from the system. In order to smooth the randomness trials to be more robust, I run the experiment ten times for each percentage values on each testbed. Table 4.9 shows the result of departure peers and their effectiveness on retrieval results on average of the ten experiments. The column entitled "Departure" refers to the percentage of peers that leave the system. The column "MAP" shows the retrieval effectiveness of the system after departing the candidate peers. The column "Relevant_Lost" corresponds to percent number of relevant documents that leave with their peers. The column "Retrieval_Degradation" shows the percent degradation in retrieval effectiveness in comparison with MAP values in Table 4.8 that shows the retrieval effectiveness without departure. As shown in two tables, the retrieval effectiveness of testbeds without replication to their corresponding testbeds with replication is significantly different on MAP metric using bootstrapping two-paired t test at $P < 0.01$.

In Table 4.9, we observe that the percent loss of relevant documents increase,

Table 4.9: The Robustness of semi-structured P2P-IR topologies Using DL* and ASIS* Testbeds.

Testbed	Departure	MAP	Relevant_Lost	Retrieval_Degradation
DLWOR	10%	0.0658	9.96%	24%
	20%	0.0655	19.29%	25%
	30%	0.0737	25.18%	15%
	40%	0.0471	45.69%	46%
	50%	0.0478	51.13%	45%
DLWR	10%	0.0221	7.15%	1%
	20%	0.0208	15.95%	7%
	30%	0.0153	34.18%	31%
	40%	0.0192	39.96%	14%
	50%	0.0164	50.31%	26%
ASISWOR	10%	0.0661	8.28%	7%
	20%	0.0603	22.54%	15%
	30%	0.0577	31.04%	19%
	40%	0.0465	41.69%	35%
	50%	0.0457	54.78%	36%
ASISWR	10%	0.0167	8.49%	3%
	20%	0.0152	16.44%	12%
	30%	0.0153	24.88%	12%
	40%	0.0134	43.42%	23%
	50%	0.0133	50.63%	23%

due to the number of the peers that depart the system (i.e. as the churn rate increases), which is not the same as the linear effect on the performance of the system. The retrieval effectiveness may have oscillatory values in different percentages. I hypothesise that this occurs due to a different number of relevant documents distributed over peers, where the probability of losing a large number of relevant documents depends on selecting the peers that have a large number of topics. If the peer contains a different number of relevant documents of multiple topics and is selected for departure, this will severely affect the overall retrieval effectiveness in the system. However, more degradation in retrieval effectiveness can be seen in testbeds without replication (i.e. DLWOR and ASISWOR) which are more affected than the testbeds with replication. The retrieval degradation differs based on the environments used, the $ASIS^*$ file-sharing system is more

robust than DL^* environments as the number of relevant documents in the DL^* environments have more chance to be lost due to the small number of peers with more documents in such environments (**RQ-4.4**). From these experiments, I can conclude that the replication in P2P settings mitigates the low effectiveness of churn rate.

4.7 Chapter Summary

I conducted experiments on a semi-structured P2P-IR network (or super-peer network) on various scales of information networks and analysed their effectiveness, efficiency, and scalability. I built the semi-structured P2P-IR network under three cluster-based topologies; K-means, Half K-means-Single-Pass, and Approximation single-Pass approaches from the cluster centroids of the peers' collections using two stages of clustering at peer and super-peer level. The analysis shows differences between topologies in retrieval effectiveness using different testbeds and as a reference I compared the results using closeness metric to the centralised systems (**RQ-4.1**). P2P networks, due to document distribution and terms' statistics, are not comparable to the centralised systems. I select a target topology (i.e, K-means) based on retrieval performance (i.e, MAP value) for further analysis. Using target K-means semi-structured P2P-IR network, a set of experimental settings was examined, such as the effect of retrieval models, the effect of increasing number of super-peers, and the scalability and robustness of semi-structured P2P-IR networks. The results show a set of conclusions as the information retrieval models have an effect in comparison to the centralised systems with semi-structured topologies, where there are retrieval models perform better in distributed systems than other models (**RQ-4.2**). The effectiveness of retrieval models comes from the size of distributed collections and the less discriminative documents in distributed systems. In terms of increasing number of super-peers, when the number of super-peers increase the retrieval effectiveness will decrease as the small number of super-peers are closer to centralised systems (**RQ-4.3**). The robustness factor that was tested for the effect of departure or failed peers shows better results as the semi-structured P2P-IR networks are imposed to be more robust and scalable (**RQ-4.4**).

Semi-structured P2P-IR networks based on the results could be an alternative framework for Information Retrieval. The discussed results depend on flooding query to all super-peers and peers in the network. The main goal of the thesis is to improve the retrieval quality through query routing. Hence I select K-means as target network based on the analysis in this chapter to enhance and propose query routing process methods to route a query to more relevant peers that contain more relevant documents as will discuss in the next two chapters.

Chapter 5

Cooperative Resource Selection Methods in Federated Search

> "The true sign of intelligence is not knowledge but imagination."
>
> — Albert Einstein, (1879-1955)

5.1 Introduction

Federated information systems provide an interface by way of a broker to traffic a given query to distributed search resources. Query routing in this environment requires a set of representations from several resources to make a decision on which of these resources contains highly relevant documents (Shokouhi and Si, 2011; Callan, 2000; Crestani and Markov, 2013; Markov and Crestani, 2014). The decision depends on the cooperation of the resources in providing information which in turn affect the quality of resource selection. As discussed in Chapter 3, federated search consists of two environments of cooperation to acquire the required representations. This chapter studies the query routing in cooperative environments in which the resources can provide statistical information about their collections. In particular, I propose cooperative resource selection methods, especially in P2P-IR and meta-search environments. In the meta-search environment, a single broker sends a query to distributed search engines, then merges the result lists of these search engines as a final merged result list to the users (Meng et al., 2002). In P2P-IR environment, there are a variety of architectures that vary in terms of the organisation of the peers. I, however, use the k-means

cluster-based architecture for evaluation (Alkhawaldeh and Jose, 2015), which is also described in Chapter 4. This cluster-based P2P architecture explores a middle-ground in the trade-off between centralised and completely decentralised P2P systems. In a nutshell, the cluster-based (or semi-structured) architecture uses two levels of clustering, which are intra-peer and inter-peer clustering as also discussed in Chapter 4. In intra-peer clustering, the documents within each peer are clustered to form lexically coherent clusters using bisecting K-means algorithm. In inter-peer clustering, the cluster centroids are collected from across peers to be clustered again in order to build the network topology using K-means algorithm into k clusters. The resultant k clusters are then handled by a separate *super-peer* as representation information of its peers denoted in Equation 4.8 on page 95.

In the 2-tier semi-structured architecture, I examine a set of cooperative resource selection methods on this homogeneous clustering groups. The study is presented in three contributions to exhibit the importance of applying several resource selection methods. The contributions are clarified as follows whereas the research questions are derived from the high-level research question HL-RQ2 in the Introduction 1 as follows:

1. I posit that simplistic word frequency based models would be able to leverage the content homogeneity in clustered P2P-IR frameworks to affect accurate resource selection. Accordingly, I adopt classical inverted indexes from IR literature for resource selection in clustered P2P-IR and propose IPI (Inverted PeerCluster Index), a remarkably simple resource selection method for clustered P2P-IR. Through an extensive empirical evaluation on classical P2P-IR testbeds, I establish that IPI matches sophisticated resource selection methods on virtually every parameter of interest.

    - **RQ-5.1:** What is the retrieval effectiveness of using the centroids of peers' clusters at the super-peer level as a resource selection method (i.e, IPI)? and how efficient is the IPI approach in terms of message complexity (number of routed peers)?

2. I hypothesise an enhanced applicability of document retrieval methods for resource selection in the context of clustered P2P-IR where routing decisions

are to be made among semantically coherent resources. I do an empirical benchmarking of document retrieval methods, against state-of-the-art resource selection methods, on the semi-structured P2P-IR architecture. The results across many testbeds establish that document retrieval methods are able to deliver consistently superior resource selection accuracies.

- **RQ-5.2:** How can we exploit the retrieval models to be used as resource selection methods in federated search environments? How effective and efficient are those methods in retrieval perspective?

3. The current classification-based state-of-the-art approaches use machine learning techniques on a training set of features built up on query logs and relevance indication values that are calculated for each query and the target resources to construct a near optimal classifier model. The model is used to predict the resources' scores for a future query to forward that query to most likely ones having relevant documents. In this chapter, I propose a Learning to Route (LTRo) approach that exploits and uses the Learning to Rank (LtR) algorithms to build a classifier in order to route a query to a set of relevant peers (or resources) in semi-structured P2P-IR networks and meta-search environments. The training set is built based on the specific resource selection methods chosen based on the effective performance of the studied resource selection methods in the contribution 2 and another two aggregated terms of query term and document frequencies. The labels in the training set are estimated based on the number of relevant documents retrieved at top 10 ranked results when traffic a query to a specific resource.

   - **RQ-5.3:** How efficient is a set of LtR algorithms to predict the testing set using IR measurement metrics? What are the retrieval effectiveness and routing efficiency of cooperative semi-structured P2P-IR networks and meta-search environments in using these approaches as resource selection?

The remainder of this chapter is organised as follows: First of all, Section 5.2 provides the testbeds used in these approaches, query routing processes, and the

experimental settings. I provide this section to explain the meta-search environment testbed as it is not used in the previous chapter as well as to use it for evaluating in document retrieval resource selection and LTRo approaches. Section 5.3 discusses the proposed approach that uses the coherent peers' clusters at the super-peer level to build an inverted index data structure of terms along with their related peers' values, to be used for routing a query to the relevant peers. In addition, I discuss the experimental results of this approach on the semi-structured P2P-IR networks. The document retrieval resource selection methods along with their experimental results are discussed in Section 5.4. Finally, Section 5.5 discusses LTRo approach and its experimental results, followed by the conclusions and discussions in Section 5.6.

5.2 Dataset and Evaluation Methodology

The evaluation scenario in semi-structured P2P-IR system is the same experimental methodology in (Alkhawaldeh and Jose, 2015), which is also explained in Chapter 4, including testbeds, queries, retrieval process, and retrieval evaluation metrics. In addition, I evaluate the document retrieval resource selection methods and LTRo approach on meta-search environments that have one broker and a set of search engines. The reason is that in meta-search environments, we have one broker who handles all the search engines. Therefore, I use the FedWeb 2013 dataset, which is a test collection with similar properties as the actual web (Demeester et al., 2013). The FedWeb 2013 data collection (13k documents, 200 queries) consists of search results from 157 web search engines in 24 categories ranging from general web search engines to small news, academic articles and images to jokes and lyrics, each of which is modelled as a separate peer. I use resource selection tasks and calculate the evaluation measures on resource selection query relevance file[1]. In meta-search systems, a broker, as a form of interface, organises and manages a set of search engines through routing the given query to relevant ones and merging the retrieved result lists into a final list.

[1] https://sites.google.com/site/trecfedweb/2013-track (October, 2016)

5.3 Inverted PeerCluster Index

The aggregated clusters at the super-peer level are considered as resources' representations for each super-peer and its peers. The representations reflect the inherent topical information of ensemble semantic peers around a super-peer. But the question is how to exploit such representation information to build an effective and efficient query routing technique in this type of semi-structured P2P-IR networks. In this section, I propose an effective and simple technique which is called Inverted PeerCluster Index (IPI) that is built from using the representation information and exhibit its experimental results. I examine the IPI approach only on semi-structured P2P-IR networks considering the clustering coherence at super-peer level. I did not construct the approach under meta-search environments because these environments consist of only one broker and a set of distributed search engines, since there are not enough super-peers to build such kind of approach.

5.3.1 Inverted PeerCluster Index Architecture

The IPI approach seeks to exploit the content coherence at the peer-cluster level in the cluster-based architecture to devise a simple scoring method that generalizes the conventional inverted indexing approach for information retrieval. The two-layered cluster-based architecture allows for resource selection at two levels: one where a subset of super-peers may be chosen using the super-peer centroids, and another where a subset of peer-clusters may be chosen at each super-peer based on the peer-cluster centroids. I will focus on the latter, assuming that the query is made available to all super-peers; this also allows for a fair comparison against single-level P2P networks.

The inverted peer index at any super-peer is simply an inverted index (i.e., word-level lists) over the peers; each peer is tagged with a score that is aggregated across the peer-clusters from the peer. The word-level index for the word w at a super-peer S_j would contain 2-tuples in the form of $[peer, score]$ entries:

$$IPI(S_j)[w] = \{\Big[P_x, \sum_{P_x^y \in C_{S_j}} Cd(P_x^y)[w]\Big] | P_x : \exists y, P_x^y \in C_{S_j}\} \tag{5.1}$$

Though the lists entries are peer-specific, I call it peer-cluster index since the corresponding scores are computed by aggregating across only those peer-clusters that belong to the super-peer (and not across *all* documents in the peer, as in other techniques). I will denote IPI much like an associative array where $L[w].P_x$ denotes the score for P_x in the word-specific list $L[w]$. It may be noted that while the $Cd(P_x^y)[w]$ scores are normalized (being an average), the $L[w].P_x$ scores are not normalized. Thus, peers that contribute more peer clusters containing w to a super-peer tend to get higher scores; this construction allows us to favor such peers in the peer ranking method. Given a query Q containing terms $\{q_1, q_2, \ldots, q_l\}$, I then score peers within S_j:

$$Score(P_x, Q) = \begin{cases} 0, & \text{if } \exists q_i,\ IPI(S_j)[q_i].P_x = \phi \\ \sum_{q_i \in Q} IPI(S_j)[w].P_x, & \text{otherwise} \end{cases} \tag{5.2}$$

Thus, only those peers who have an entry in the list corresponding to *each* query term are *eligible* assuming that all query terms are important to reflect the information need of users; the eligible peers are then scored using simply a sum-based aggregation of corresponding entries. Though this eligibility condition introduces a discontinuity in the scoring function, it is computationally attractive since peers can be discarded as soon as it is discovered that they do not figure in any one of the l query-term specific lists.

Depending on the budget constraint (in terms of a number of peers to choose), the peer-clusters with the top scores are chosen for S_j to route the query to. Declaratively, if k peer-clusters are to be chosen,

$$\textit{Top-}k@(S_j, Q) = \underset{R \subseteq C_{S_j}, |R|=k}{\arg\max} \sum_{P_x \in R} Score(P_x, Q) \tag{5.3}$$

The typical resource allocation algorithm chooses k as a specified fraction of *eligible* peers according to the selection mechanism adopted by that approach. The fraction operates as a meta-parameter to the selection algorithm. The selected peers, then process the query in a cluster-agnostic manner.

Example: The IPI is a data structure of posting lists that connects a dictionary of terms (vocabulary or lexicon) with a list of postings. The IPI simulates the inverted index in traditional information retrieval models with some difference as the inverted index of traditional systems use documents where the proposed

model IPI uses peers. Since, each term in the dictionary has a list of postings where each posting contains peers along with approximate term score. The term score in this example is cheap to obtain and is calculated by Equation 5.1. The list of term and their postings called posting list, which can be shown in Figure 5.1.

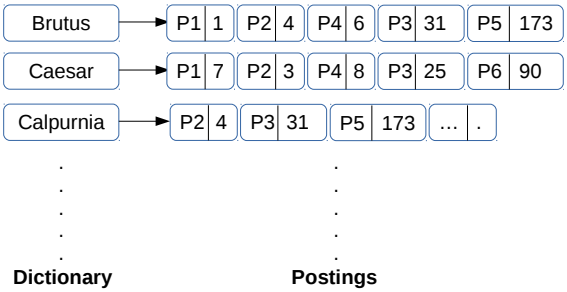


Figure 5.1: Inverted Peer Index.

The figure shows a map between a term and a set of postings. This structure will increase the performance of search as the IPI data structure is a hash map of terms and their posting list. The super-peer looks for a specific term in $O(1)$ to retrieve the posting list of that term. The intersection approach for merging posting lists of query terms takes $O(m + n)$ where m represents the number of terms and n represents the number of postings for a specific query term.

In the semi-structured P2P network, the query processing occurs at the super-peers level, the super-peer uses the IPI data structure to select the largest relevant peers for the given query. If the query has one term, the super-peer selects the postings of that term and decide how many peers should be selected for the query processing (i.e, k in Equation 5.3). On the other hand, if the query has more than one term, the super-peer intersects the posting lists of the query terms by summing the terms' scores of intersected peers together. Then, the super-peer sort the final result in descending order to select more relevant peers that have the whole query terms as shown in following examples.

Case 1: If the query has one term, for example **Caesar**, the super peer takes the postings list of term Caesar and sorts the list in descending order

based on the weight of the term in the postings. From the Figure 5.1 the final result for peer selection is as follows:

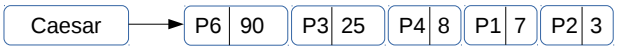


Figure 5.2: An Example of IPI Data structure (Caesar).

Case 2: If the query has more than one term for example **Brutus and Calpurnia**. The super peer intersects the postings lists of two terms simultaneously and sums the terms' scores of common peers between the two lists. In Figure 5.1, the two lists for the two terms are merged and the final result is sorted in descending order as follows:

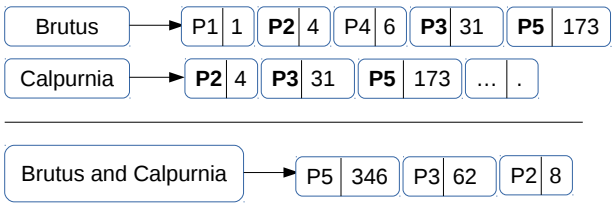


Figure 5.3: An Example of IPI Data Structure (Brutus and Calpurnia).

P_2, P_3, and P_5 are eligible since they occur in two lists of the terms while the other peers are not.

5.3.2 Inverted PeerCluster Index: Experimental Results

While I presented the IPI routing approach and clarified the importance of exploiting the clusters at the super-peer level as a resource selection method, this subsection discusses the experimental results for examining its effectiveness and efficiency. The results of the empirical study over six testbeds and 10% of the resource selectivity parameter are listed in Table 5.1. Two core dimensions are of interest; effectiveness as measured by the quality of results on IR evaluation metrics, and the efficiency as measured by the messaging costs indicated by the number of peers chosen. The table lists Precision, Recall, P@{10, 30, 100} and

MAP measures evaluated at the top-1000 in the merged list, which are the core parameters of interest in large-scale P2P-IR. The results of IPI approach is compared against the state-of-the-art cooperative resource selection approaches described in more detail in Chapter 3: CVV (Yuwono and Lee, 1997); Taily (Aly et al., 2013); CORI (Callan et al., 1995; Callan, 2000); KL Xu and Croft (1999); vGIOSS (Gravano et al., 1999); in addition to RW (Random Walk) method.

The number of selected peers at the super-peer level chosen at 10% for each testbed while there are other five percentages used in evaluation appears in Appendix A.1.

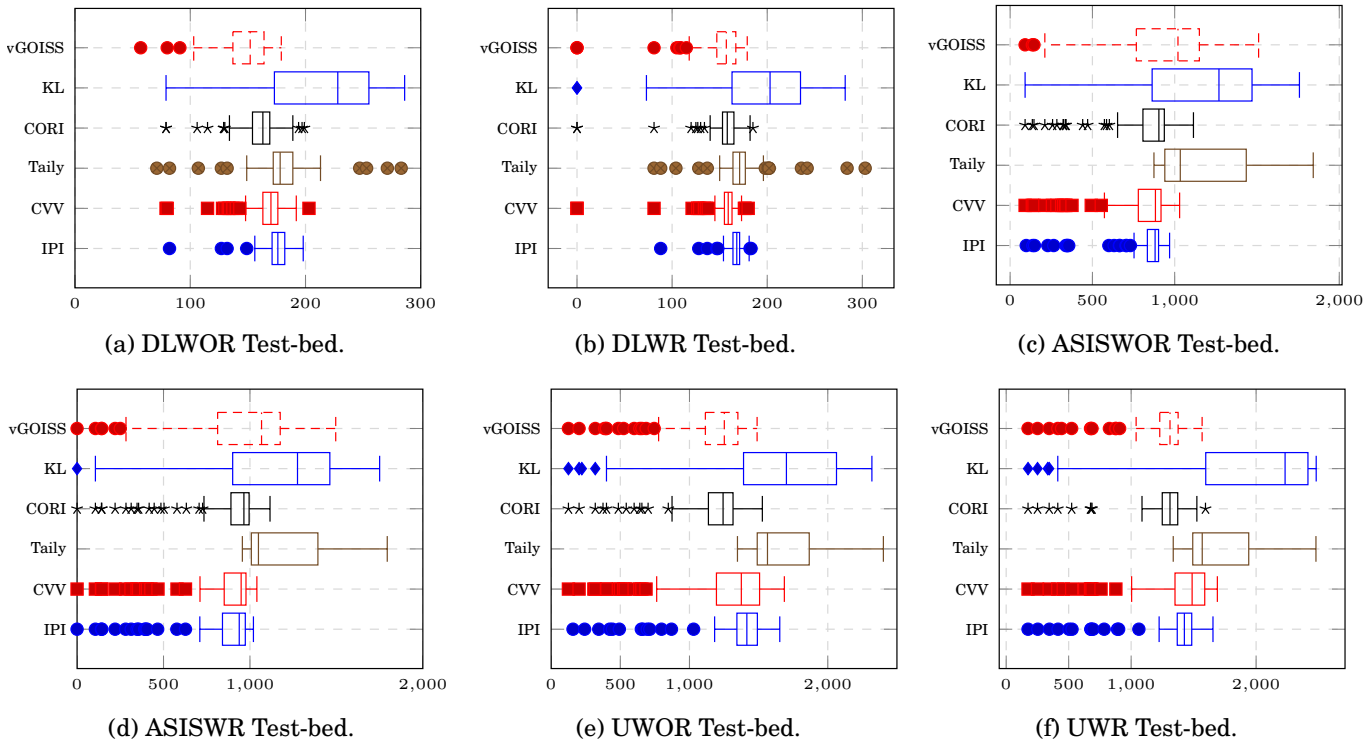
Table 5.1: IPI Retrieval effectiveness at 10% of Selected Peers ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t-test).

DL*	**DLWOR Testbed**						**DLWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
Flooding	0.02866	0.54790	0.16900	0.13233	0.08770	0.08659	0.02089	0.42564	0.01837	0.01837	0.02306	0.02232
IPI	**0.02461**	**0.47601**	**0.188**$^{\bullet}$	**0.135**	**0.084**	**0.09358**	**0.0229**	**0.4534**	**0.023**	**0.02266**	**0.0294**	**0.02662**
CVV	0.02435	0.44472	0.182	0.13133	0.0825	0.08281	0.02158	0.41428	0.02	0.021	0.0301	0.0239
Taily	**0.02769**	**0.50463**	**0.182**	**0.137**	**0.0898**	**0.09493**	**0.02523**	**0.46911**	**0.026**	**0.025**	**0.0331**	**0.02978**
CORI	**0.02686**	0.45697	0.177	**0.139**	**0.0878**	0.0864	**0.0256**	0.44471	**0.027**	**0.03366**	**0.0399**	**0.03171**
KL	0.01105	0.1513	0.109	0.07233	0.0427	0.02782	0.0104	0.13508	**0.034**	**0.038**	**0.0322**	0.01295
vGIOSS	0.02208	0.38807	0.171	0.12266	0.077	0.07616	0.0183	0.33898	**0.034**	**0.04533**	**0.051**	**0.02915**
RW	0.01455	0.18186	0.134	0.093	0.0509	0.03856	0.01516	0.13561	**0.042**	**0.03633**	0.028	0.0122
IPI Rank	3	2	1	3	3	2	3	2	6	6	6	4
ASIS*	**ASISWOR Testbed**						**ASISWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
Flooding	0.02532	0.46296	0.16400	0.11966	0.07620	0.07122	0.01635	0.33847	0.01414	0.01818	0.01899	0.01732
IPI	**0.02469**	**0.45859**	**0.163**	**0.11766**	**0.077**	**0.07047**	**0.02117**$^{\bullet}$	**0.41612**$^{\circ}$	**0.016**	**0.01967**	**0.0217**	**0.01998**
CVV	**0.02508**	0.45601	**0.166**	**0.12233**	**0.0776**	**0.07227**	0.02053	**0.41176**	**0.016**	0.01933	0.0215	0.01954
Taily	**0.02613**	**0.45867**	0.159	**0.12233**	**0.0785**	**0.07103**	**0.02088**	0.39249	**0.02**	**0.02133**	**0.0233**	**0.02218**
CORI	**0.02636**	**0.46426**	0.161	**0.122**	**0.0785**	**0.0743**	0.02066	0.39906	0.014	0.019	**0.0228**	**0.02014**
KL	0.01525	0.25003	0.121	0.08666	0.0534	0.04275	0.01565	0.27106	0.015	**0.02067**	**0.0296**	0.01721
vGIOSS	0.01966	0.35336	0.153	0.106	0.0664	0.06321	0.01901	0.35926	**0.016**	**0.02667**	**0.0347**	**0.02386**
RW	0.0131	0.16287	0.111	0.06933	0.0407	0.03037	0.01451	0.17353	**0.019**	**0.02167**	0.0204	0.01084
IPI Rank	4	3	2	4	4	4	1	1	3	5	5	4
U*	**UWOR Testbed**						**UWR Testbed**					
Flooding	0.02764	0.49910	0.21200	0.14800	0.09580	0.10581	0.02269	0.42657	0.01400	0.01900	0.02450	0.02331
IPI	**0.02582**	**0.46564**	**0.211**$^{\bullet}$	**0.15267**$^{\bullet}$	**0.0951**$^{\circ}$	**0.10022**	**0.02269**	**0.42657**	**0.014**	**0.019**	**0.0245**	**0.02331**
CVV	0.02158	0.38	0.186	0.13466	0.0785	0.0862	0.01932	0.36159	0.012	0.01867	0.0224	0.01926
Taily	**0.02781**	**0.48616**	0.183	**0.14633**	**0.0932**	**0.10931**	**0.02563**	**0.45538**	**0.014**	**0.02067**	**0.0292**	**0.02751**
CORI	**0.0273**	**0.46807**	**0.191**	0.14466	0.0928	0.09365	**0.02674**	**0.46429**	**0.014**	**0.02066**	**0.0331**	**0.02969**
KL	0.01295	0.20201	0.122	0.083	0.0438	0.0478	0.01319	0.24118	**0.015**	**0.02067**	**0.0249**	0.01491
vGIOSS	0.01859	0.34657	0.157	0.114	0.0655	0.06733	0.01906	0.35583	**0.016**	**0.028**	**0.0405**	**0.02367**
RW	0.01405	0.16086	0.117	0.08233	0.0439	0.03062	0.01463	0.17359	**0.02**	**0.023**	0.0181	0.00832
IPI Rank	3	3	1	1	1	2	3	3	4	6	5	4

Table 5.1 shows the retrieval results of the IPI approach on the three environments of two testbeds where the boldface values refer to the IPI values as well as the other methods' values that have competitive results. The underlined cases describe the best values of the IPI approach compared to the best boldface baseline method. The IPI rank listed under each scenario-metric combination summarizes the competitiveness of IPI in comparison to the listed resource selection methods except the Flooding method. Flooding method routes the query to all the peers under a specific super-peer, and hence have better performance, and it is used here only for reference. In general, as shown, the simple IPI approach performs on par with state-of-the-art resource selection methods, often outperforming them.

In terms of retrieval effectiveness, IPI approach is seen to be better than a majority of the resource selection techniques under study especially the testbeds without replication. In particular, the average ranks of IPI approach on the DLWOR and DLWR testbeds are 2 and 5 respectively. The retrieval results in the ASIS* environment are slightly low with high ranks of 4 and 3 on average for the ASISWOR and ASISWR testbeds respectively. In the uniform environments, IPI approach has the same average rank as the DLWOR testbed with respect to the UWOR and 4 on average using the UWR testbed. The poor results in testbeds with replication (i.e, the high IPI rank values), however, might reflect the effect of overwhelming the peers with relevant documents of different topics that affects the distribution of peers in the system. Hence, this replication deviates routing decision to these peers that might have a small number of relevant documents and at the same time contain more non-relevant documents to the topic of given query. In addition, we can see that the effective result of the IPI approach on the ASISWOR testbed declines on average rank of 4 in comparison with the DLWOR and UWOR testbeds. This result might be due to the small size of peers causes an effect on the clustering quality results in their documents ends up with poor clusters at super-peer level. In contrast, the ASISWR testbed obtains better results more than the DLWR and UWR testbeds presumably due to the replicated relevant documents that are concentrated more in specific peers of small size.

In terms of the efficiency aspect of the IPI approach, Figure 5.4 shows boxplots of message complexity on the testbeds over hundred queries (i,e. 451-550



(a) DLWOR Test-bed. (b) DLWR Test-bed. (c) ASISWOR Test-bed.

(d) ASISWR Test-bed. (e) UWOR Test-bed. (f) UWR Test-bed.

Figure 5.4: The Efficiency of IPI Approach in Semi-structured P2P-IR Network.

topics). In the DL* environment, the IPI approach has less message complexity with values between 150 and 200 of routed peers. The average and standard deviation values for DLWOR and DLWR are (175, 14) and (166, 11) respectively. The performance is much better in the ASIS* environment on average of 811 and 836 of messages with a standard deviation of 191 and 237 for ASISWOR and ASISWR respectively in comparison with the baseline methods. This improvement in efficiency is due to the small number of peers that are grouped around the super-peers' clusters where the query is routed to fewer coherent and topical peers. The IPI approach performs differently in the U* environment, as shown in Subfigures 4.4(e) and 4.4(f), the routed peers in this approach are condensed between 1300 and 17500. Specifically, on average performance, the UWOR testbed has 1326 of routed peers with 315 standard deviations whilst the UWR testbed obtains 1337 of routed peers on average under 307 standard deviations. The reason behind high number of messages in the DL* and U* environments appears to be due to the relatively large number of documents in their peers, which increases

the number of inter-peer clusters with more chance to be at multiple super-peers for the selection decision.

In summary, the IPI approach exploits the coherent clusters created from multiple peers to help routing a query at super-peer level to the peers that are most likely to contain relevant documents. This means the IPI approach uses fewer messages reducing network traffic and at the same time it has comparable, retrieval effectiveness values as discussed before. The IPI approach is a very competitive resource selection method, on par with the state-of-the-art, for the clustered P2P-IR framework at super-peer level.

5.4 Document Retrieval Methods for Resource Selection

In document processing techniques, the system considers its documents to be having a few topics and discriminates these documents up on the topics they have. This scheme is not applicable in general P2P networks if we assume that each peer is as a big document of documents (or collection). This is due to the absence of content coherence at the super-peer level which leads to that the resource selection layer needs to work across diverse resources. For example, in a typical case for federated search over various news agencies, each news agency managed by a separate peer would comprise documents as diverse as the entire corpus. This property of general P2P-IR has led to the development of methods that model and exploit distributional information (e.g., variance) of terms across peers, in their scoring process (Yuwono and Lee, 1997; Xu and Croft, 1999; Aly et al., 2013). Much like in traditional IR, P2P-IR resource selection works by computing a query-specific score for each peer, followed by choosing the top-scoring peers to route the query to. However, we can see this coherence by exploiting the cluster hypothesis in the clustered semi-structured P2P-IR networks and the document retrieval models could be employed as resource selection methods. In this section, I explain the idea behind using the conventional retrieval models in IR as resource selection methods in semi-structured P2P-IR networks and meta-search environments. I also discuss the retrieval effectiveness and routing efficiency to examine their applicability and performance.

5.4.1 Document Retrieval Methods Assumptions

In general P2P-IR without any intra-peer clustering, different resources at the resource selection level could potentially have widely varying content; this makes it infeasible to consider these resources as analogous to the documents, since document processing techniques often are built with the assumption that only a few topics are touched upon, in each document. An example is the concentration parameter in LDA that enforces topic sparsity at the document level (Wei and Croft, 2006). As another example, two large documents comprising text segments from across domains could achieve a high tf-idf cosine value due to a lot of small similarities adding up and end up being comparative with a pair of documents from the same domain; thus, tf-idf cosine is better applied in cases where documents exhibit a good amount of lexical skew (akin to topical focus). In short, the absence of content coherence in general P2P-IR architectures has made it infeasible to exploit the advances in document processing directly for resource selection, resulting in a divergent evolution of techniques for the tasks. The central hypothesis is that the clustered P2P architecture, by virtue of clustering, helps bring back the analogy between documents and resources, thus recalling document relevance methods into contention for resource selection (Alkhawaldeh et al., 2016).

The document retrieval methods that I use in the benchmarking study are TF-IDF, Okapi.BM25, Hiemstra language model (LM), BB2, In_expB2, In_expC2, InL2, and DFI0 model. The methods vary from different families which are discussed in more detail in Chapter 2. None of the used techniques, with the exception of TF-IDF (Melucci and Poggiani, 2007), have been studied for resource selection in hierarchical P2P-IR. Since these techniques are specialized to rank documents (i.e., sets of words), resources under each super-peer needs to be modelled as documents. I do this by simply using the *big document* model. Under this model, one big document is created for each resource managed by a super-peer; as a toy example, $SP2$ in Figure 4.1 would be searching for two documents, one built by collating the documents it manages from P3 and the other formed by the document subset of P2. Having defined the big documents, resource selection is just about ranking the big documents using the chosen model and routing the

query to the resources corresponding to the top-ranked big documents.

Example: BB2 resource selection method is a Divergence From Randomness (DFR) method, as discussed in Chapter 2 in Equation 2.30, that has term-weight inversely related to the probability of term-frequency within the resource r obtained by a model M of randomness. Therefore, the informative importance of a term to a resource can be estimated through examining how much the term frequency distribution departs from a random distribution of terms' frequencies in that resource. Correspondingly to resource selection method, the BB2 model is a Bose-Einstein distribution for randomness as the following equation:

$$Score(t|r) = \frac{tf_R + 1}{tf_{t,r} \cdot (tfn + 1)}\big(- \log_2(R-1) - \log_2(e) + f(R + tf_R - 1, R + tf_R - tfn - 2) - f(tf_R, tf_R - tfn)\big) \quad (5.4)$$

where R is the number of resources under a specific broker. tf_R is the term frequency of t under a specific broker. tf_r is the resource frequency of t. tf_n is the normalised term frequency. It is given by the normalisation 5.5:

$$tf_n = tf_{t,r} \cdot \log_2(1 + c \cdot \frac{avgrl}{rl}) \quad (5.5)$$

where c is a free parameter (in this case c = 1) and the relation f is given by the Stirling formula as:

$$f(n, m) = (m + 0.5)log_2(n/m) + (n - m)log_2 n \quad (5.6)$$

Table 5.2: Adapted Resource Selection Methods Terms Terminology

Document (d)	Resource (r)	Description
$tf_{t,d}$	$tf_{t,r}$	term frequency in d or r
dl	rl	document or resource length in number of terms
$avgdl$	$avgrl$	Average document or resource length in number of terms
$df_{t,c}$	rf_t	document or resource frequency of t term
N	R	number of documents or resources
tf_C	tf_R	term frequency in whole collection or in whole resources
N_t	R_t	number of unique terms in Collection or Resources

Table 5.2 contains the converted terminologies of retrieval models on documents to be used as resource selection method in federated web search. These terminologies applied to other adapted resource selection methods as well.

Table 5.3: Document Retrieval Methods effectiveness and efficiency at 10% of Selected Peers ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t-test compared with the best baseline method).

DL*	**DLWOR Testbed**							**DLWR Testbed**						
Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
CVV	166	0.02435	0.44472	**0.182**	0.13133	0.0825	0.08281	155	0.02158	0.41428	0.02	0.021	0.0301	0.0239
Taily	180	**0.02769**	**0.50463**	**0.182**	0.137	**0.0898**	**0.09493**	173	0.02523	**0.46911**	0.026	0.025	0.0331	0.02978
CORI	160	0.02686	0.45697	0.177	**0.139**	0.0878	0.08639	155	**0.02559**	0.44471	0.027	0.03366	0.0399	**0.03171**
KL	215	0.01105	0.1513	0.109	0.07233	0.0427	0.02782	198	0.0104	0.13508	0.034	0.038	0.0322	0.01295
vGIOSS	**148**	0.02208	0.38807	0.171	0.12266	0.077	0.07616	**151**	0.0183	0.33898	0.034	**0.04533**	**0.051**	0.02915
RW	306	0.01455	0.18186	0.134	0.093	0.0509	0.03856	244	0.01516	0.13561	**0.042**	0.03633	0.028	0.01219
BB2	162	**0.02893**$^\circ$	**0.50619**	**0.182**	**0.151**$^\bullet$	**0.0951**$^\bullet$	**0.10129**	156	**0.02674**	**0.47434**	0.025	0.03033	0.0415	**0.03336**
In_expB2	161	**0.02853**	0.4869	0.172	**0.145**	**0.092**	0.09242	155	**0.0261**	0.45713	0.026	0.02799	0.0402	0.03136
In_expC2	161	**0.0285**	0.48757	0.178	**0.1467**	**0.0925**	0.09487	155	**0.02588**	0.45612	0.026	0.02766	0.0394	0.03109
InL2	159	**0.02888**$^\circ$	0.49467	**0.184**	**0.1477**$^\circ$	**0.0939**$^\circ$	0.09317	153	**0.02762**$^\bullet$	0.46483	0.031	0.036	0.0481	**0.03643**$^\bullet$
Hiemstra_LM	162	0.02605	0.44801	**0.192**	**0.14**	**0.0907**	0.08853	153	**0.02571**	0.42881	**0.04**	**0.059**$^\bullet$	**0.067**$^\bullet$	**0.04098**$^\bullet$
DFI0	158	0.0273	0.46549	**0.204**	**0.144**	**0.0916**	0.09311	**150**	**0.02673**	0.44828	0.031	**0.0457**	**0.056**	**0.03869**$^\circ$
TF_IDF	159	**0.02909**$^\bullet$	0.49665	**0.185**	**0.15**$^\bullet$	**0.0963**$^\bullet$	0.09467	154	**0.028**$^\bullet$	**0.4761**	0.031	0.03533	0.0475	**0.03644**$^\bullet$
BM25	**130**	0.01615	0.23338	0.102	0.073	0.0512	0.05280	**145**	0.01617	0.27057	0.027	0.02767	0.0323	0.02480
Majority	1	5	1	5	7	7	1	2	7	2	-	2	2	5
ASIS*	**ASISWOR Testbed**							**ASISWR Testbed**						
Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
CVV	**793**	0.02508	0.45601	**0.166**	**0.12233**	0.0776	0.07227	**847**	0.02053	**0.41176**	0.016	0.01933	0.0215	0.01954
Taily	1190	0.02613	0.45867	0.159	**0.12233**	**0.0785**	0.07103	1205	**0.02088**	0.39249	**0.02**	0.02133	0.02330	0.02218
CORI	819	**0.02636**	**0.46426**	0.161	0.12200	**0.0785**	**0.0743**	868	0.02066	0.39906	0.014	0.019	0.0228	0.02014
KL	1130	0.01525	0.25003	0.121	0.08666	0.0534	0.04275	1147	0.01565	0.27106	0.015	0.02067	0.02960	0.01721
vGIOSS	926	0.01966	0.35336	0.153	0.106	0.06640	0.06321	957	0.01901	0.35926	0.016	0.02667	**0.0347**	**0.02386**
RW	1003	0.0131	0.16287	0.111	0.06933	0.0407	0.03037	1072	0.01451	0.17353	0.019	**0.02167**	0.0204	0.01084
BB2	821	**0.02689**	**0.47238**$^\circ$	**0.176**$^\bullet$	**0.132**$^\bullet$	**0.0812**	**0.07838**$^\bullet$	871	**0.02103**	0.403	0.015	0.02	0.0233	0.02079
In_expB2	822	0.02593	0.45701	**0.171**	**0.132**$^\circ$	**0.081**	**0.07837**$^\circ$	870	**0.02204**	0.41	0.015	0.02033	0.0238	0.02206
In_expC2	821	0.02583	0.45321	**0.174**$^\circ$	**0.133**$^\circ$	**0.0804**	**0.07852**$^\circ$	868	**0.0224**	0.40898	0.015	0.02066	0.0244	0.02235
InL2	832	**0.02708**$^\circ$	0.46345	**0.172**	**0.13**	**0.082**$^\circ$	**0.077**$^\bullet$	882	**0.0215**	0.40954	0.015	0.01933	0.023	0.02098
Hiemstra_LM	937	0.02577	0.43893	**0.166**	**0.127**	**0.0813**	**0.07434**	962	**0.0252**$^\bullet$	**0.44**$^\circ$	0.018	**0.025**$^\circ$	**0.036**	**0.02815**$^\bullet$
DFI0	**742**	0.01472	0.11529	0.097	0.06	0.03570	0.01597	**852**	**0.0246**$^\circ$	**0.438**$^\bullet$	**0.019**	**0.0237**	0.0293	**0.02615**$^\circ$
TF_IDF	1153	0.01549	0.25344	0.12	0.086	0.0536	0.0427	874	**0.02155**	0.40747	0.015	0.01933	0.02320	0.02087
BM25	872	0.02509	0.44041	0.16	0.117	0.0766	0.07216	916	**0.02134**	**0.4152**	0.014	0.01833	0.0223	0.02058
Majority	1	2	1	5	5	5	5	-	8	3	-	2	1	2
U*	**UWOR Testbed**							**UWR Testbed**						
Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
CVV	1257	0.02158	0.38	0.186	0.13466	0.0785	0.0862	1384	0.01932	0.36159	0.012	0.01867	0.0224	0.01926
Taily	1707	**0.02781**	**0.48616**	0.183	**0.146**	**0.0932**	**0.10931**	1730	0.02563	0.45538	0.014	0.02067	0.0292	0.02751
CORI	1163	0.0273	0.46807	**0.191**	0.14466	0.0928	0.09365	1257	**0.02674**	**0.46429**	0.014	0.02066	0.0331	**0.02969**
KL	1609	0.01295	0.20201	0.122	0.083	0.0438	0.04780	1972	0.01319	0.24118	0.015	0.02067	0.02490	0.01491
vGIOSS	**1154**	0.01859	0.34657	0.157	0.114	0.0655	0.06733	**1231**	0.01906	0.35583	0.016	**0.028**	**0.0405**	0.02367
RW	2071	0.01405	0.16086	0.117	0.08233	0.0439	0.03062	2425	0.01463	0.17359	**0.02**	0.023	0.0181	0.00832
BB2	1195	**0.0297**$^\circ$	**0.5**$^\circ$	**0.214**$^\bullet$	**0.159**$^\bullet$	**0.1024**$^\bullet$	0.10671	1304	**0.0282**$^\circ$	**0.511**$^\bullet$	0.017	0.02266	0.0321	**0.0322**$^\circ$
In_expB2	1189	0.02459	0.42607	**0.204**	0.144	0.0876	0.08779	1299	0.02378	0.43229	**0.021**	0.02766	0.03520	0.02752
In_expC2	1186	0.02392	0.41939	**0.199**	0.1417	0.0848	0.0892	1294	0.02338	0.42865	0.018	**0.02833**	0.03610	0.02725
InL2	1161	**0.0309**$^\bullet$	**0.5276**$^\bullet$	**0.206**	**0.148**	**0.1021**$^\bullet$	**0.11066**	1254	**0.03**$^\bullet$	**0.533**$^\bullet$	0.019	0.02266	0.036	**0.0344**$^\bullet$
Hiemstra_LM	**1137**	**0.03**$^\circ$	**0.5054**	**0.212**$^\circ$	**0.155**$^\circ$	**0.1013**$^\bullet$	0.10671	**1225**	**0.0302**$^\bullet$	**0.521**$^\bullet$	**0.021**	**0.02833**	**0.0435**	**0.037**$^\bullet$
DFI0	**1151**	**0.0285**	**0.4886**	**0.217**$^\bullet$	**0.156**$^\circ$	**0.1**	0.106	1252	**0.0277**	**0.5056**$^\bullet$	**0.021**	**0.028**	0.0403	**0.034**$^\bullet$
TF_IDF	1167	**0.031**$^\bullet$	**0.5353**$^\bullet$	**0.21**	**0.153**	**0.104**$^\bullet$	0.10887	1259	**0.0299**$^\bullet$	**0.5377**$^\bullet$	0.017	0.022	0.0354	**0.0337**$^\bullet$
BM25	**1135**	0.0216	0.38975	0.161	0.121	0.0766	0.09792	**1130**	0.02075	0.3827	**0.024**	0.025	0.0352	**0.0302**
Majority	3	5	5	7	5	5	1	2	5	5	4	3	1	6

5.4.2 Experimental Results

Through experimental results of using document retrieval methods as resource selection approaches, I investigate the applicability of deploying the conventional IR retrieval models as resource selection methods over semi-structured P2P-IR networks. Table 5.3 summarizes the results on the IR testbeds. The top and bottom parts of the table illustrate standard resource selection methods and the document retrieval respectively; the best value of each evaluation metric is boldfaced for each category of techniques. Additionally, the best value across categories is underlined as well. It may be seen from the table that the document retrieval methods perform better in every testbed and evaluation metric, barring one case where Taily is seen to outperform others. The last row Majority counts the number of document retrieval methods (out of the 8 being studied) that perform either equally or better than the best performing general resource selection method under study, on the corresponding evaluation metric. Appendix A.2 includes the other five majority percentages as deep evaluation of these methods.

As shown, the document retrieval methods exhibit best and competitive results as well as a significant performance at specific metrics, as the best performing resource selection methods, compared to the standard baseline methods in three environments. The average majority percentages on each testbed are 54.2%, 37.5%, 48%, 33.3%, 58.3%, and 50% to DLWOR, DLWR, ASISWOR, ASISWR, UWOR, and UWR respectively; which is on average over all environments roughly 47% across metrics. In more detail, the probabilistic TF_IDF and BM25 models rank the peers based on how rare are the query terms to discriminate the relevance of peers through tuning the TF and IDF (i.e, ICF inverted collection Frequency) within the peer and other peers. These two methods perform better in DL and U environments due to the large size of peers in determining the two variables. On the other hand, Hiemstra_LM method obtains consistent results in the three environments, which indicates the effectiveness of relevant peers' (i.e, the whole peers under specific super-peer) language models to generate the query terms especially the testbeds with replication contain more relevant documents. The DFR methods, showing another view of relevance based on statistical distributions, explore more important results as the randomness considered the key

evidence in examining the relevancy. I believe that the randomness in sharing the documents in the peers is steadily higher as the users have full control over their sharing information in P2P scenario. However, specifically in this case, this kind of methods could have the ability to identify the randomness of query terms over the peers and select those with a high score of informativeness in the occurrence of query terms. The results show much better performance of DFR methods with even more competitive values compared to other methods. In the same line of relation to DFR models, DFI0 method distinguishes the ratio of query terms' occurrences to be independent of the peers and select those peers having more distinctive and contributed terms. In other words, the relevant peers contain the query terms with a different constant ratio of terms frequency rather than the other peers are selected to be routed at super-peer level. This approach maintains a competitive performance to the baseline methods with best results at specific metrics, as the relevant peers are typically condensed around specific super-peers. This means that the terms of a specific topic corresponding to their topical peers have different frequency ratio and are not independence in these peers compared to the others.

In respect of routing efficiency, the documents retrieval methods achieve competitive results in reduced numbers of routing peers in approximately 90% of the cases especially to CORI, Taily, and CVV methods. In brief, BM25 method, specifically in DL and U environments, has best routing efficiency compared to the vGOISS as best baseline method while DFI method maintains best value in ASISWOR and competitive value in ASISWR compared to CVV method.

Table 5.4 summarizes the corresponding results on the federated web search testbed, in the same format as for P2P-IR testbeds. The results on this testbed also confirm the superior performance of the document retrieval methods for resource selection in the meta-search environment of federated search with significant results over all metrics excepts BM25 approach. It may also be noted that five of eight document retrieval methods significantly outperform the best performing general resource selection method, on each analysed metric, which is Taily approach.

In summary and based on the results, the clustered P2P architectures, by virtue of clustering, help bring back the analogy between documents and re-

Table 5.4: Document Resource Selection Retrieval Results on FedWeb2013 Testbed ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t-test compared with the Taily method).

Method	Precision	Recall	P@10	MAP	nDCG@20
CVV	0.05072	0.473	0.32535	0.16042	0.18397
Taily	**0.05795**	**0.54193**	**0.38083**	**0.22278**	**0.218**
CORI	0.04392	0.40285	0.29653	0.14143	0.15167
KL	0.0191	0.1730	0.1604	0.0445	0.1071
vGIOSS	0.0347	0.3220	0.2431	0.1072	0.1318
RW	0.0171	0.1497	0.1556	0.0401	0.0874
BB2	**0.0654**$^\bullet$	**0.61735**$^\bullet$	**0.4523**$^\bullet$	**0.2665**$^\bullet$	**0.271**$^\bullet$
In_expB2	**0.0635**$^\bullet$	**0.5957**$^\bullet$	**0.4203**$^\bullet$	**0.2542**$^\bullet$	**0.251**$^\bullet$
In_expC2	**0.0635**$^\bullet$	**0.5957**$^\bullet$	**0.42**$^\bullet$	**0.2542**$^\bullet$	**0.251**$^\bullet$
InL2	**0.0628**$^\bullet$	**0.5811**$^\bullet$	**0.429**$^\bullet$	**0.251**$^\circ$	**0.252**$^\bullet$
Hiemstra_LM	0.0485	0.4347	0.3090	0.1583	0.1766
DFI0	0.0447	0.4030	0.3285	0.1563	0.1903
TF_IDF	**0.0639**$^\bullet$	**0.5980**$^\bullet$	**0.4375**$^\bullet$	**0.258**$^\bullet$	**0.256**$^\bullet$
BM25	0.0179	0.2242	0.1865	0.1164	0.1123
Majority	5	5	5	5	5

sources. This establishes that document retrieval methods are very effective for resource selection in the clustered P2P-IR architecture and should be preferred against classical resource selection methods designed for general P2P-IR. From the federated search perspective, we can see also the effectiveness of using document retrieval models as resource selection. These results answers the research question **RQ-5.2**.

5.5 LTRo: A Learning to Route Approach

In order to examine the classification-based resource selection approaches in the semi-structured P2P-IR networks and meta-search environments, I use Linear Regression (LR) (Theil, 1992) and Multi-Layer Perception (MLP) (Chen and Manry, 1993) learned classifier to rank and score the resources (or peers) to route a query to the most relevant ones. In addition, I proposed a Learning To Route (LTRo) approach that uses LtR algorithm discussed in Chapter 2 to build a classifier for the same task as classification-based approaches. The training set of

used classifiers is built up on the document resource selection methods where I use the most effective approaches of different families as features and aggregated features along with label assigned on top 10 relevant retrieved documents. This section explains the LTRo approach and the experimental results using the semi-structured P2P-IR networks and meta-search environments.

5.5.1 LTRo: Assumption and Architecture

LTRo approach trains a supervised learner on a set of evidence from different peers to rank and route the queries to expectedly the most relevant ones. Each supervised learning algorithm requires a training set of feature vectors along with their labels to create a classifier. In particular, specifically in learning to rank algorithms and in the scope of resource selection, the training set consists of a group-based query of related resources' features along with assigned labels. Formally, $q_i(i = 1, \cdots, n)$ corresponds to the set of n queries for the training step and $r^{(i)} = \{r_1^{(i)}, \cdots, r_{m(i)}^{(i)}\}$, with m number of resources associated with query q_i as feature vectors. The labels are formulated as $y_i(i = 1, \cdots, n)$, which is a set of relevance judgements for each query q_i and feature vector $r^{(i)}$.

Preprocessing phase. In the experiments, I generate a training set of features and labels using 10,000 training query selected from 1.6 million known-item queries[1], which are 18.82% (one term), 47% (two terms), 19.7% (three terms), 10.57% (four terms), 2.75% (five terms), and 1.16% (six terms) to mimic real-life scenarios. The used features are six single resource selection methods and other two aggregated terms. The single evidence methods are CORI (Callan et al., 1995), CVV (Yuwono and Lee, 1997), and high quality adapted document retrieval methods such as BB2, Hiemstra_LM, DFI0, and TF-IDF as shown in experimental results earlier. The other two features are aggregated document and term frequency of the query terms.

Arguello et al. (2009a) and Hong et al. (2010) proposed methods to assign relevance label to a specific resource. The method in (Arguello et al., 2009a) traffics a query to a centralised full dataset index where the resource is considered to be relevant if more than $\tau = 3$ of its documents are contained in the retrieved result list. Since this method is not feasible in real-life environments, Hong et al.

[1] http://boston.lti.cs.cmu.edu/callan/Data/P2P (October, 2016)

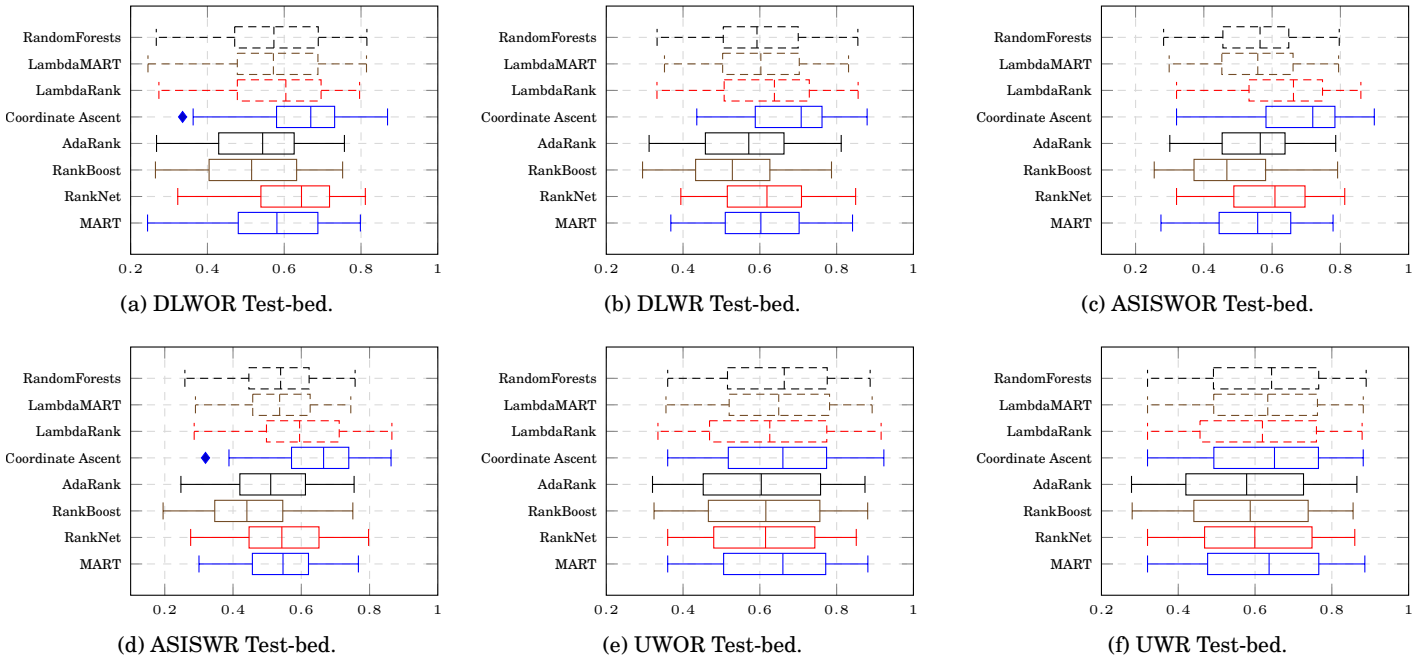
(2010)'s method routes the query to each resource and the number of relevant retrieved documents when it is greater than a specific threshold τ ($\tau = 3$ for 100 average number of relevant documents and $\tau = 1$ otherwise) determines the relevance of the resource. Since the second method is more effective and realistic, I follow their approach in creating the relevance labels for the resources. I assign a numerical label, which is the number of relevant retrieved documents within the top 10 ranked document list for each resource. These numerical labels are suitable in learning to rank algorithms that optimise evaluation metrics.

Training phase. I use a set of LtR approaches[1] to build various ranking learners to study their effectiveness as resource selection methods. The feature vectors are normalised using the z-score method (based on mean and standard deviation) (Jain et al., 2005). For optimising the training data, I use the ERR@10 measurement (Chapelle et al., 2009). Furthermore, for validation, I use a standard 5-fold cross validation approach where the results are averaged. However, in the semi-structured P2P network, I build a model for each super-peer using the training set created over its peers.

Testing phase. The learned ranking model in each super-peer is applied to a new given query in order to sort its peers according to their relevance scores. The ranked peer set is then routed for a given query, with selected threshold k determining the percentages of peers to be routed. In this paper, I used six values of k: 5%, then 10% to 50% in steps of 10%. Effectiveness values were averaged over all these percentages.

Baseline approaches. LR and MLP (feedforward artificial neural network) baseline approaches are used as classification-based approaches on the training set using weka[2]. In addition, I compare the LTRo approach with the six single resource selection methods that are used for creating the training set and the Taily approach as a recent cooperative state-of-the-art resource selection algorithm (Aly et al., 2013).

[1] https://sourceforge.net/p/lemur/wiki/RankLib/ (October, 2016)
[2] http://www.cs.waikato.ac.nz/ml/weka/ (October, 2016)



(a) DLWOR Test-bed. (b) DLWR Test-bed. (c) ASISWOR Test-bed.

(d) ASISWR Test-bed. (e) UWOR Test-bed. (f) UWR Test-bed.

Figure 5.5: The Prediction of LTRo approaches on Testing set

5.5.2 LTRo: Experimental Results

I study the retrieval effectiveness and the efficiency of using the LTRo approach on the testing query set. In terms of retrieval quality, I examine the effectiveness of two different methods, one explains the accuracy of the LTRo approach in predicting the testing set labels while the other method focuses on evaluating LTRo approach using retrieval results of using TREC query topics (i.e, 451-550) conducted under a semi-structured P2P-IR network.

In terms of predictive testing set, Figure 5.5 shows boxplots of predictive results related to the testbeds of the three environments. X-axis represents nDCG@10 values predicted for queries in the testing set, whilst y-axis represents the LTRo approaches. As shown, Coordinate Ascent approach outperforms the other approaches with competitive results in U environment. On average over all approaches, the values roughly are between 0.4 and 0.7 in three environments, which indicates the accuracy of LTRo approaches in peer selection with superior effectiveness to Coordinate Ascent approach. In order to examine the retrieval effectiveness of the LTRo approaches, I conduct experiments on semi-structured

P2P-IR network through routing the testing queries and evaluate the retrieval effectiveness on the final merged result list. At first, classification-based resource selection and LTRo approaches significantly outperform the single baseline and Taily methods on all measurement metrics in the two testbeds of each environment as shown in Table A.7 on page 208. The retrieval effectiveness is estimated under two-paired statistically significant bootstrap t-test with $p \leq 0.01$, but I do not put a significant sign as it is significant over all metrics. In comparing the LTRo approach with the state-of-the-art classification-based approaches, Table 5.5 exhibits the retrieval results at evaluation metrics where LTRo-LR and LTRo-MLP refer to the percent improvement to the two classification approaches. Since the LTRo approaches have the same retrieval results, I combine the average results as a LTRo name in Table 5.5. Each LTRo approach assigns different scores to the same peers with relative orders. The LTRo approach obtains competitive results in almost all the metrics with improvements in most of the cases.

Table 5.5: LTRo Retrieval effectiveness.

DL*	**DLWOR testbed**						**DLWR testbed**					
Method	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
LR	0.04017	**0.57933**	0.22633	0.17783	**0.11153**	0.13322	**0.03687**	**0.54805**	0.05650	0.05561	0.05142	0.06389
MLP	**0.04028**	0.57889	0.22683	**0.17833**	0.11148	0.13135	0.03679	0.54706	0.05650	**0.05572**	0.05160	0.06389
LTRo	0.03972	0.56756	**0.23015**	0.17731	0.11045	**0.13470**	0.03668	0.53493	**0.05767**	0.05539	**0.05258**	**0.06506**
LTRo-LR	**(-3.77%)**	**(-16.32%)**	**(+7.48%)**	**(-1.29%)**	**(-5.1%)**	**(+3.46%)**	**(-1.57%)**	**(-15.22%)**	**(+3.71%)**	**(-0.79%)**	**(+6.85%)**	**(+3.47%)**
LTRo-MLP	**(-4.8%)**	**(-18.77%)**	**(+6.04%)**	**(-2.55%)**	**(-5.12%)**	**(+7.55%)**	**(-0.94%)**	**(-14.23%)**	**(+3.71%)**	**(-1.19%)**	**(+5.71%)**	**(+3.48%)**
ASIS*	**ASISWOR testbed**						**ASISWR testbed**					
LR	0.04356	0.54821	0.24500	0.17606	0.10615	0.12134	**0.03965**	0.52124	**0.06400**	0.05900	0.04972	0.06027
MLP	0.04354	0.54891	0.24400	0.17561	0.10598	0.12099	0.03954	0.52066	**0.06400**	0.05900	0.04965	0.06021
LTRo	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
LTRo-LR	**(+2.76%)**	**(+6.73%)**	**(+1.07%)**	**(+1.98%)**	**(+0.73%)**	**(+5.34%)**	**(-0.28%)**	**(+2.46%)**	**(-1.72%)**	**(+3%)**	**(+3.3%)**	**(+4.18%)**
LTRo-MLP	**(+2.85%)**	**(+6.27%)**	**(+2.16%)**	**(+2.69%)**	**(+1.2%)**	**(+5.98%)**	**(+0.25%)**	**(+2.89%)**	**(-1.72%)**	**(+3%)**	**(+3.56%)**	**(+4.33%)**
U*	**UWOR testbed**						**UWR testbed**					
LR	0.08842	0.73835	0.47133	0.35867	0.21113	0.32898	0.08488	0.71121	0.12317	0.12383	0.09630	0.14347
MLP	0.08842	0.73908	0.47233	0.35861	0.21110	0.32908	0.08485	0.71135	0.12317	0.12394	0.09628	0.14342
LTRo	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
LTRo-LR	**(+0.24%)**	**(+2.33%)**	**(+0.98%)**	**(+0.88%)**	**(+1%)**	**(+1.7%)**	**(+0.17%)**	**(+0.12%)**	**(+4.31%)**	**(+1.83%)**	**(+4.77%)**	**(+2.71%)**
LTRo-MLP	**(+0.23%)**	**(+2.03%)**	**(+0.61%)**	**(+0.91%)**	**(+1.02%)**	**(+1.66%)**	**(+0.22%)**	**(+0.07%)**	**(+4.31%)**	**(+1.72%)**	**(+4.8%)**	**(+2.76%)**

In the meta-search shown in Table 5.6, the LTRo approach has better and competitive results with almost all the cases compared to the single-based baseline methods except Taily method at Precision, Recall, and MAP values as well as the BB2 method in all the measurement metrics. In comparison to classification-based approaches, the LTRo approach has significant improvements on average of 5.5%.

Figure 5.6 explains the message complexity (or efficiency) of the LTRo ap-

Table 5.6: LTRo Retrieval Results on FedWeb2013 Testbed.

Method	Precision	Recall	P@10	MAP	nDCG@20
Taily	0.05795	0.54193	0.38083	0.22278	0.21750
CORI	0.04392	0.40285	0.29653	0.14143	0.15167
CVV	0.05072	0.47300	0.32535	0.16042	0.18397
BB2	**0.06545**	**0.61735**	**0.45230**	**0.26648**	**0.27078**
Hiemstra_LM	0.04845	0.43465	0.30902	0.15828	0.17660
DFI0	0.04472	0.40295	0.32847	0.15625	0.19032
TF_IDF	0.06392	0.59797	0.43750	0.25785	0.25562
LR	0.05250	0.49330	0.39165	0.20768	0.21622
MLP	0.05250	0.49330	0.39165	0.20768	0.21622
LTRo	**0.05560**	**0.51813**	**0.41285**	**0.22120**	**0.22845**
LTRo-LR + MLP	(+**5.9%**)	(+**5.03%**)	(+**5.41%**)	(+**6.51%**)	(+**5.7%**)

proach and LR, MLP, and Taily baseline methods, which is calculated as a number of routed peers. LTRo approach achieves a consistent significant performance with a small number of routed peers compared to Taily approach and better performance to LR and MLP. This reduces the number of messages in the network with quick response to a given query. In particular, LTRo approach has fewer number of routed peers with approximately 3.9% and 38.3% on average for the classification-based approaches and the Taily method respectively in DL^* environment on both testbeds. In $ASIS^*$ environment on both testbeds, the LTRo approach obtains a reduction in a number of routed peers with approximately 0.5% and 88.5% on average for classification-based and Taily methods receptively. Finally, LTRo approach gains high performance in message complexity with approximately 0.95% and 80.75% on average for classification-based and Taily approaches in U^* environment and its testbeds. Consequently, these efficiency results satisfy the assumption that LTRo approach is preferable than the other baseline methods.

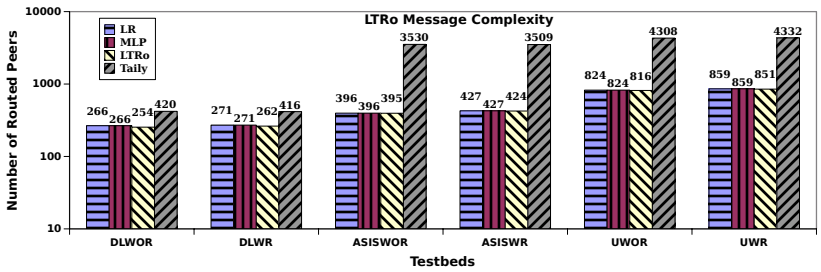


Figure 5.6: LTRo efficiency on number of selected Peers on Three environments.

In summary, the LTRo approach naturally matches the performance of the learning to rank approaches in the information retrieval field, which theoretically might be a preferable choice for resource selection in federated search. The results indicate the applicability of the assumption in using LtR approaches to improve resource selection and lead to answer the research question **RQ-5.3**.

5.6 Conclusions and Discussions

In this chapter, I discussed the problem of query routing in a cluster-based semi-structured P2P-IR architecture. This architecture emphasises the idea of grouping the similar semantic peers together into a centrally managed node called super-peer using clustering algorithms. I used two clustering steps exclusively, which are in-peer and out-peer clustering, to identify the topics in each peer and then build the network through combining the peers with similar topics around specific super-peer. However, under cluster-based network, I clarified three query routing hypothesises; which are exploiting the clusters at the super-peer level as peers' representations for the routing decision, enhancing the applicability of using the traditional document retrieval methods as routing technique, and creating a classification-based approach built over a training set generated under such network using specific features.

In more details, I outlined a simple index and method for resource selection, called IPI, extended the conventional inverted index in the IR to the level of peer-clusters. Through an extensive analysis on P2P-IR testbeds, IPI, is seen to rival state-of-the-art resource selection methods designed for general P2P-IR architecture, both in terms of accuracy as well as messaging costs; this establishes IPI as a simple and effective resource selection method for clustered P2P-IR and subsequently answer the research question **RQ-5.1** that evaluates the possibility of using the centroids of peers' clusters for resource selection at super-peer level. Back to the document retrieval methods track, I empirically benchmarked well-known document retrieval methods against the state-of-the-art resource selection methods designed for general P2P-IR. An extensive analysis of retrieval effectiveness over P2P-IR using classical IR evaluation metrics validate the hypothesis convincingly, with document retrieval methods consistently outperform-

ing the others. This establishes that document retrieval methods ought to be the preferred choice for resource selection in clustered P2P-IR environments and answers the research question **RQ-5.2** that examines exploiting the document retrieval models as resource selection approaches. On the third track of this study, I proposed a Learning to Route (LTRo) approach in the semi-structured P2P-IR network to traffic the given query to most likely relevant peers. I used the state-of-the-art LtR algorithms to train a model and predict the peers' scores for unknown queries. The training set is built over specific features of resource selection approaches from the former study while the label of feature vectors is assigned as the number of the relevant documents at top 10 retrieved documents. By experimenting under different conditions, over multiple testbeds, I studied the performance variations of the LTRo approach. The results show the effectiveness of the LTRo approach on testing set and an improvement in the routing process which consequently confirm the answer of research question **RQ-5.3** in what the effectiveness and efficiency using LtR algorithms as resource selection methods. Finally, I discussed the IPI, document retrieval and LTRo methods and analysed their applicability on the K-means architecture of semi-structured P2P-IR network. These methods are different in the representation of resources and the process of ranking to route the query to most relevant resources. In spite of the effectiveness and the efficiency of the discussed methods, they do not exploit the users to improve the retrieval quality. Since, and due to the importance of interfering the users to improving the results, I will explain how to exploit the implicit feedback of the users on the retrieved results to enhance the retrieval quality of the system in the next Chapter 6.

Part III

Query Routing Using Implicit User Feedback

I highlight the reputation-based query routing approaches in more details. This includes simulating user behaviour in providing implicit feedback and the process of building a reputation-based data structure for query routing. Furthermore, I study different scenarios reflecting real-life user behaviours in P2P-networks and how these have an impact on the quality of routing a query to relevant peers.

Chapter 6

Reputation-based Query Routing

"The way to gain a good reputation is to endeavor to be what you desire to appear."

— Socrates, (469-399 B.C)

6.1 Introduction

In Peer-to-Peer Information Retrieval (P2P-IR) scenarios, the reputation values could be exploited to improve retrieval effectiveness as in online commercial systems. The reputation values are often extracted from past user ratings for supporting users' decision-making. The basic idea in online commercial systems is to allow parties to rate each other and using the aggregated ratings to assist other users in deciding whether or not to transact with that party in the future (Jøsang et al., 2007). Variations of such approaches are exploited in e-commerce systems, collaborative systems, and security aspects of the P2P systems. Regarding security aspect of the P2P networks, the systems retrieve, filter, and evaluate the recommendations using past behaviour of the individual peers to assign a high reputation score to a specific peer, for example, one that provided more trustworthy documents in the past (Zhang, 2011). Thus these systems enhance the security features through the "punishment" of those untrustworthy peers in the network and this helps to reduce the security risk for the newly joined peers (Jøsang et al., 2007). However, reputation scores are not exploited to support effective query routing. I propose to exploit these scores to route the given query to most reputable peers, which provided more relevant documents in the past.

In a position paper, Kazai and Milic-Fralying proposed a trust-based retrieval approach for a social network where actors (users) and documents are included to form a data graph (Kazai and Milic-Frayling, 2008). They conjectured that it is possible to exploit users' feedback (voting technique like reviews on a product) to build a reputation rank of actors in the system and use them for retrieval.

The idea is to exploit users' behaviours, especially the implicit feedback a user naturally gives when using a P2P retrieval system. Such actions/feedbacks include viewing a document, clicking a document, and/or downloading a document. However, prior research has ignored users' feedback on documents for enhancing the retrieval effectiveness in P2P-IR systems. The behaviour of users in P2P networks has been exploited only to improve the security aspects of P2P systems (Jøsang et al., 2007). Hence I develop a reputation-based query routing approach in semi-structured P2P-IR network by mining the reputation measures from past interaction data. The proposed approach essentially monitors implicit users' interactions on clicking or downloading relevant documents from the retrieved ranked list to build reputation scores for the participating peers as relevant indicators. These actions of accessing a peer or downloading a document have shown power-law patterns in the past (Saroiu et al., 2002), which means in real-world P2P systems there is a high dependency on a set of peers or documents. The proposal is to exploit this power-law pattern of reputation, which will increase the retrieval effectiveness and at the same time decreases the number of messages in P2P networks as will be discussed in this chapter that addresses the high level research question **HL-RQ3**.

HL-RQ3: *Is implicit feedback provided by the users during their interactions in semi-structured P2P-IR networks effective for improving query routing, and how should such feedback be exploited to build reputation data structures?*

The remainder of this chapter is organised as follows: Section 6.2 discusses the reputation concept from abstraction and relevance perspectives, explains the reputation-based approaches exploiting the past user interaction data, and a simulation of user's behaviour in providing implicit feedback through clicking and downloading documents. Section 6.3 explains the evaluation methodology used to validate the proposed reputation-based approaches estimated through the re-

trieval effectiveness and efficiency (message complexity). Section 6.4 studies the retrieval effectiveness of the proposed approach as well as the network efficiency on a number of messages by comparing the results against the CORI and Taily resource selection approaches. Section 6.5 discusses the limitations of using the reputation-based approaches as query routing methods in semi-structured P2P-IR networks, followed by the conclusions in Section 6.6.

6.2 Reputation-based search in Semi-structured P2P-IR network

In this section, I first discuss "*reputation*" as a psychological and social concept. Subsequently, I clarify the concept of reputation as an indicator of relevance to both documents and their peers through simulating the user interactive behaviour. This section also presents proposed reputation-based routing approaches and discusses their techniques to find the most reputable and relevant peers.

6.2.1 Reputation concept

Reputation is quite a complicated concept recurring in our daily lives in various forms, and can be viewed as a summarization and abstraction of human brain directly and indirectly formed by their opinions of past, current or future actions. According to Abdul-Rahman and Hailes (2000), reputation can be defined as an expected perception about an individual's behaviour created directly (personal experience) or reported indirectly by others (recommendations or third party verification) through the past history of interactions. In online communities, reputation is derived from the underlying sources as an indication of the trustworthiness of a source to interact with (Jøsang et al., 2007). Reputation does not necessary mean trustworthiness as we can trust a non-reputable person through direct interactions and the untrusted person could be reputable through indirect interactions reported by others. In social network systems, the reputation values are spread over the network and can transfer the parties' reputation values from one community context to another and provide an agreement between different community context to exchange reputation values (Ruohomaa and Kutvonen, 2005). The challenge is to redefine and adapt the concept of reputation for effec-

tive and efficient retrieval in P2P-IR systems.

6.2.2 Reputation and Relevance

The proposed approach is based on mining users' interaction (implicit feedback) during past searches. During a search process, a user sends a query and from the returned result list selects relevant documents, which could potentially come from different peers. The approach monitors and then mines these interaction data to assign reputation values to documents and peers. I consider the reputation as what is generally said or believed about a person's or thing's character or standing (Abdul-Rahman and Hailes, 2000). Here, reputability does not necessarily mean popularity, as reputability has to be estimated (or rated) by a group of users while popularity is just seen to be known (not rated) by the users (Li, 2012). Hence, the reputation value of a document can be defined as the number of times this document has been clicked or downloaded in the past searches. In real-life P2P scenarios, users might click or download non-relevant documents in the search process, and hence the reputation is not just restricted to the relevant documents. Although non-relevant documents could also be reputable, just from collective interactions, it is expected that the relevant documents get assigned more values of reputation than the non-relevant documents. In other words, the relevant documents obtain a high probability to be clicked or downloaded by the users more than the non-relevant documents. The reputation value of a document is estimated using Equation 6.1.

$$Rep(d_k) = \sum_{i=1}^{nf(d_k)} f(P_i, d_k), \tag{6.1}$$

where $nf(d_k)$ refers to the total amount of feedback for document d_k received from different peers and $f(P_i, d_k)$ denotes the feedback for document d_k received from peer P_i.

In the semi-structured P2P-IR network, the super-peers in the system manage and organise the reputation values of their peers for resource selection and result merging processes. The reputation values of peers, however, are calculated by aggregating its documents' reputation values. Technically, a super-peer S_j build a hash map ($P_{s_j} = \{\vec{P_{s_j}1}, \vec{P_{s_j}2}, ..., \vec{P_{s_j}n}\}$) of their n peers and documents reputation

vector as $\vec{P_{s_ji}} \equiv P_{s_ji} \Rightarrow (Rep(d_1), Rep(d_2), ..., Rep(d_n))$, where P_{s_ji} is peer i belong to super-peer S_j and $Rep(d_k)$ refers to document reputation value of document d_k as calculated in Equation 6.1. The peer's reputation value is aggregated from the other users' feedback on its documents as in Equation 6.2.

$$Rep(P_{s_ji}) = \sum_{k=1}^{nd(P_{s_ji})} Rep(d_k), \tag{6.2}$$

where $nd(P_{s_ji})$ denotes the total number of documents for peer P_{s_ji} at super-peer S_j. This is an aggregated score from the reputation scores of its documents. Thus, if a peer has a number of highly reputable documents, then its reputation value will be high. The reputation value of a peer reflects the probability of finding relevant documents in that peer; the higher the reputation value of a peer, the higher the probability of locating relevant documents. In the past, Saroiu et al. (2002) showed that, in real-life P2P systems, a power-law distribution pattern of usage is demonstrated; which means a high dependency on a set of peers and a set of documents. This could be due to the implicit reputation values that are frequently assigned by users on past reliable interactions on these rated documents and then on their peers. In addition, the reputation of peers gets spread through collective actions or spreading of information about the availability of the information. The proposed approach on assigning reputation values reflects this behaviour and I propose to use these values in routing queries.

Figure 6.1, as an example, exhibits a semi-structured P2P-IR network of 50 super-peers (SP) and their peers (P) as well as a reputation data structure on super-peer SP_j and a user that interacts with a retrieval ranked list of submitted query, say, "universities in Scotland". In particular, the user sends the query "universities in Scotland" to the semi-structured P2P-IR network and receives a merged ranked list of documents' links (e.g, "University of Glasgow"). The user at this time downloads or clicks the links of relevant documents as shown by the tick sign. The implicit relevant information of this interaction or session is sent back automatically to the super-peers that are responsible for the clicked or downloaded documents of their peers. The super-peer SP_1, as shown in Figure 6.1, builds a reputation data structure of peers and their rated documents; for example P_1 has $Doc_4, Doc_{10}, ...$, and Doc_{50} with 17, 5, ..., and 1 amount of

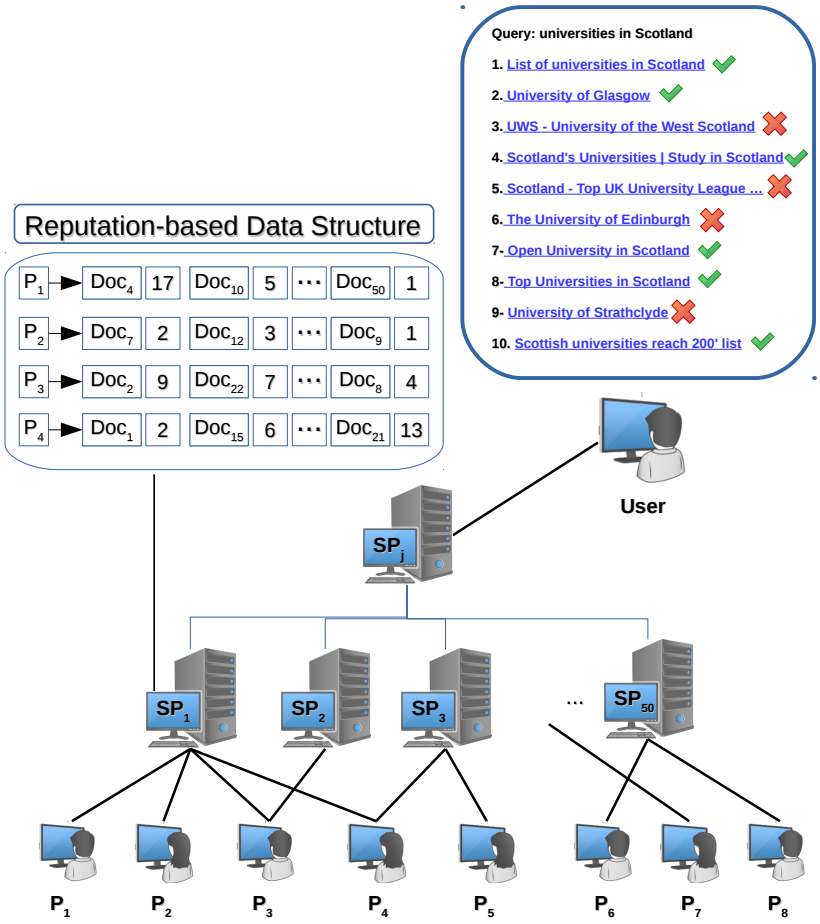


Figure 6.1: Simulating User Data Interaction

feedback, respectively. The reputation value of peer P_1, based on this usage information is aggregated over these scored values to be more than or equal to 23 as I do not know the reputation values of other documents on this example. At query run time, the super-peer SP_1 ranks their peers for a new query on different techniques, as I will discuss in Subsection 6.2.3, and sends the query to highly relevant reputable peers. In addition, such a selection process can be guided by content relevance as well.

6.2.3 Reputation-based Query Routing Methods

Once the proposed approach mines the implicit interaction data, each super-peer manage a data structure of peers and their documents' reputation values. This

data structure guides the super-peers to determine the reputable peers to route the query. I propose four reputation-based routing methods to select the most reputable peers for a given query. These methods are derived and formulated depending on previous work in reputation-based systems (see Section 3.4.5 on page 78) through clicking and/or downloading documents as reputation measures for security purposes. The proposed approaches are in more detail as follows:

1. **Reputation-based selection method (R).** In this method, each super-peer routes a given query to its peers based on their reputation values estimated by Equation 6.2. The super-peers route the query to peers with high reputation values. However, this does not guarantee relevance for the current query because the reputation values are generated from past interactions potentially of different query topics. Although the R method ignores the content-based information between the query and the peers, it is robust and can select highly reputable peers that most likely contain relevant documents based on past users' interaction data.

2. **Prioritised Reputation method (RP).** This method uses the **R** approach first to select those peers that are highly reputable and most likely to contain relevant documents to the given query. The returned documents of the selected peers are merged at the super-peer level to form a combined list. In the RP method, this is guided by the reputation values of the reputable documents in the returned list and ranks them on the top of the result list. The reputable documents' scores retrieved by the information retrieval model are increased by aggregating the reputation scores of the documents to that value. Each peer in P2P-IR systems might potentially have different retrieval models and hence have variation in document scores. The result merging algorithm at super-peer level normalises the document scores of each peer result list and then merges these lists. In the evaluation process, CombMNZ merging algorithm in Equation 4.6 on page 90 aggregates the document's reputation score to its normalised retrieved value. RP method re-values this list with document reputation values. Both R and RP methods are query-independent. In the following, I propose to use content features in routing.

3. **Reputation and Term statistics method (RT).** This selection approach combines the reputation values and the query terms' statistical information at the super-peer level to rank the peers. First, peers are ranked based on the reputation values using **R** method (Equation 6.2). The term statistical information of a peer is counted using Equation 6.3.

$$Cont(P_{s_j i}) = \sum_{q_t \in Q}^{|Q|} \sum_{k=1}^{nd(P_{s_j i})} TF(q_t, d_k), \tag{6.3}$$

where $TF(q_t, d_k)$ denotes the term frequency of the term q_t in reputable document d_k. This aggregates the query term occurrences in the reputable documents of a peer. The combinational score of the peer to the given query is estimated as follows:

$$Score(P_{s_j i}) = P(Q/P_{s_j i}) = \alpha * Rep(P_{s_j i}) + \beta * Cont(P_{s_j i}) \tag{6.4}$$

where α and β (or 1-α) are two positive parameters whose values determine the relative importance of reputation versus content statistics for the given query. I select the α and β parameter values experimentally as 0.4 and 0.6 respectively. In addition, I propose a method that uses the CORI score as in Equation 3.3 to be used as content, statistical information in Equation 6.4 instead of $Cont(P_{s_j i})$ information, which is called RCORI approach. CORI score of a peer summarises the importance of query terms in the peer represented by document frequency and the peers (or collection) frequency under super-peer level. This score is inspired from *tf.idf* weight of the Information Retrieval systems considering peers as a big collection of documents. In RCORI approach, the term statistical information of a peer in Equation 6.4 is replaced as follows:

$$Score(P_{s_j i}) = P(Q/P_{s_j i}) = \alpha * Rep(P_{s_j i}) + \beta * CORI_{score}(P_{s_j i}) \tag{6.5}$$

4. **Prioritised Reputation and Term statistics method (RPT).** In this method, I use **RT** approach first to select the peers, then similar to **RP** method, a super-peer re-ranks the document scores in the merged result

list through incorporating the reputation values and their content statistical information (Equation 6.3) into normalised scores. A parameter setting is assigned to exhibit the importance of reputation value over the content, statistical information in promoting the document. This merging approach prioritises the documents of high reputation values and at the same time with much content to the query terms to the top of the final result list. In addition, I use the CORI score as content, statistical information instead of just simple term statistics, which is called RPCORI approach. The RPCORI approach uses RCORI approach first to select the peers and then promotes the reputable documents in the same way as RPT approach using the content statistics of query terms' frequencies on the reputable documents.

The above methods use the peers' reputation vectors to rank the resources by selecting the highly reputable and/or relevant peers for query routing. In contrast, the resource selection algorithms use query terms' statistical information to rank and select those relevant peers for query routing. Hence I use the following state-of-the-art algorithms as the baseline methods, which are CORI (Callan et al., 1995) and Taily (Aly et al., 2013) resource selection methods.

6.2.4 Simulating User Interaction Data

The idea of reputation-based approach to P2P-IR systems is to mine users' past interaction data to assign reputation scores to documents and peers. However, in the evaluation testbeds, I do not have any past user interaction data. Hence, I simulate user interaction and assume that the user may click a relevant document if he/she is presented with a ranked list. I also consider a scenario where users may click some non-relevant documents. The simulation is conducted under three steps as follows:

Preprocessing phase (or Training queries): I use the TREC topics 451-550 to generate 100 simulated queries for each topic (or $Y = \{y_1, y_2, ..., y_{100}\}$) to be used as training set and use the corresponding query relevance file as ground truth to simulate user interactions. It is a two step process to generate the queries as

explained below:

1. I run the original queries on the centralised WT10g documents index using DPH retrieval model[1]. Then, I selected the top 10 documents for each query (or topic) as pseudo-relevant documents to extract the candidate query terms. The candidate terms represented as a set of 2-tuples in the form of [term, score] (e.g, $t_A = [t_A^{term}, t_A^{score}]$) entries as follows:

$$T_{D_y} = \{[t, \sum_{i=1}^{10} tf \cdot idf(t, d_i^y)] | \; \forall t \in d_i^y \wedge \; d_i^y \in D_y\}. \tag{6.6}$$

where D_y represents the top 10 documents of the query topic y as $D_y = \{D_1^y, ..., D_{10}^y\}$ from which the terms are extracted. $tf \cdot idf(t, d_i^y)$ is the $tf \cdot idf$ of the term t in retrieved document d_i^y from the set D_y.

2. The set of extracted terms along with their values from previous process (i.e., T_{D_y}) were used to generate queries for simulation (training queries) by combining them together (phrases of two terms as the average query term is 2.23 words as discussed in Subsection 4.2.3). The most likely phrases were selected by using a co-occurrence method called Tanimoto co-occurrence function (Pérez-Agüera and Araujo, 2008) as follows:

$$rel(q, t_A \wedge t_B) = \sum_{t_C \in q} (t_A^{Score} + t_B^{Score}) * Tanimoto_{D_y}(t_A \wedge t_B, t_C) \quad \forall t_A, t_B \in T_{D_y} \wedge t_A \neq t_B \tag{6.7}$$

t_A and t_B are two tuples of terms and values as discussed before; $(t_A^{term}, t_A^{Score})$ and $(t_B^{term}, t_B^{Score})$ respectively.

$$Tanimoto_{D_y}(t_A \wedge t_B, t_C) = \frac{c_{ABC}}{c_C + c_A + c_B - c_{ABC}} \tag{6.8}$$

c_{ABC} refers to the number of times that the three terms t_A^{term}, t_B^{term}, and t_C^{term} occur together where c_A, c_B, and c_C refers to the number of times the terms occur separately in top 10 documents, respectively.

[1]The DPH retrieval model is selected as a retrieval model because it has been shown as a better retrieval effectiveness on average at P@10 metric in comparison with other retrieval models (Wilkie and Azzopardi, 2014).

The top 100 training queries ($Q^{y}_{training}$) for each topic y are selected from the top relevance values of the phrases on Equation 6.7 using following heuristic rules:

(a) Terms that begin with numbers are eliminated.

(b) If the two top-ranked terms based unigram appear to be a phrase in the top-ranked phrases based on the bigram, these two terms are replaced by these phrases.

Training phase: The system, in this phase, issues the 10,000 queries (i.e. 100 for each TREC topic) randomly from different peers for each query. Then, the system follows the flooding approach in semi-structured P2P-IR models to route the query to the peers and super-peers as shown in Figure 4.2. Once the sender receives the final result list, the simulation algorithm mimics the behaviour of users as shown in Figure 6.1: if a user is involved, he/she would have downloaded or clicked the relevant documents in the final ranked result list. The proposed approach uses the assessment judgement file of TREC topic 451-550 to determine the relevant documents in the result list. Consequently, the system selects 10% or less of relevant documents randomly from the result list, assuming that the user implicitly downloads or clicks them as occurred in a real-life user behaviour in P2P-IR systems. The feedback results of relevant documents are sent back to their super-peers that manage the peers containing such documents. Each super-peer create and/or updates the reputation values of the downloaded or clicked relevant documents contained in their peers using the Equation 6.1. In addition, to simulate the user behaviour on relevant documents, I assume that the user clicks and downloads non-relevant documents occasionally. Hence in the simulation, I incorporate such interaction information.

Testing phase: In the evaluation protocol, original topics 451-551 are used for the evaluation. In the testing phase, however, the system follows four scenarios to validate the reputation-based query routing approaches as I will discuss in evaluation methodology Section 6.3. At each super-peer, the peers are ranked based on resource selection strategies (that is, CORI, Taily, R, RP, RT, RCORI, RPT, and RPCORI methods). After peer ranking, the super-peer routes the

query to a proportion of peers. The percentage used for peer selection are 10%, 20%, 30%, 40% and 50% of ranked peers.

6.3 Evaluation Methodology

6.3.1 Retrieval Effectiveness

I evaluate the reputation-based query routing approaches by computing the peers' reputation scores on different scenarios. These scenarios reflect how robust the approaches are in ranking and selecting the most relevant peers. The scenarios are as follows:

- **Scenario 1**: The reputation scores of the peers at each super-peer are estimated on over all training interaction data. In more details, assume we have a set of query topics $Y = \{y_1, y_2, ..., y_n\}$. Each query topic y has a set of training queries ($Q^{y}_{training} = \{q_{y1}, q_{y2}, ..., q_{y100}\}$) as computed in Subsection 6.2.4; which are 100 training queries. The super-peers, in the training phase, compute the documents' reputation values based on the training queries of query topics. The reputation of documents can be formulated on various topics; for example, the reputation score of a document d_k estimated on the amount of feedback from a few peers using queries of topic y_i is specifically estimated as follows:

$$Rep(d_k, y_i) = \sum_{q_{yi} \in Q^{yi}_{training}} f(P_{s_j i}, d_k, q_{yi}), \tag{6.9}$$

  where $Rep(d_k, y_i)$ refers to the reputation score of a document d_k using the training queries of topic y_i such as query q_{yi}. $f(P_{s_j i}, d_k, q_{yi})$ denotes the feedback of peer $P_{s_j i}$ for the document d_k on the merged result list of the query q_{yi} that associated with the topic y_i. In this scenario, once the testing query topic y_i, that is used for generating training query $Q^{y_i}_{Training}$, arrives, a super-peer S_j estimates the reputation scores of their peers as follows:

$$Rep(P_{s_j i}) = \sum_{y_j \in Y} Rep(d_k, y_j), \ \forall \ d_k \in \vec{P_{s_j i}} \tag{6.10}$$

where $\vec{P_{s_j i}}$ denotes the vector of reputable documents along with their reputation values as illustrated in Figure 6.1 that are related to the peer $P_{s_j i}$. In this scenario, the reputation scores of the peers depend on the whole interaction data of the topics, which is simply as estimated in Equation 6.2.

- **Scenario 2**: The reputation scores of the peers using scenario 1 include the interaction data from the training topics of the query. This makes scenario 1, although it simulates the real-life scenario of P2P networks, to be biased due to using the interaction data of training queries that are generated by the given query topic. In order to solve this problem, I suggest another scenario for evaluation, which excludes the interaction data of the training queries related to the testing query topic from the estimation of reputation scores. In this "leave-one-out" method, I exclude training topics derived from a given testing query, when considering the interaction data. Leave-one-out is one of cross validation methods used to evaluate a hypothesis model during training phase (Witten et al., 2011). In the cross validation methods, the training set is divided into k subsets (or k-folds) where a set of folds are used as training set while the rest is used as validation set. Given such division technique, the training phase is conducted under k repeated times. Each time, one of the k subsets is used as validation set and the other *k-1* subsets are aggregated to form a training set. Then the average error across all k times is estimated to evaluate the model. In the leave-one-out method, "one" subset is selected as validation set while the other sets are combined as training set at each k trials (Vehtari et al., 2016). Hence, here in scenario 2, I use such technique in different way through excluding the interaction data of training queries that are generated by given testing query from estimating the reputation score of a specific peer. In scenario 3, on the same way, a variation of excluding interaction data is used on three k-fold percentages. In a formal way, the reputation score of a peer $P_{s_j i}$ in leave-one-out method is estimated at super-peer level as follows:

$$Rep(P_{s_j i}, y_i) = \sum_{y_j \in Y \wedge y_j \neq y_i} Rep(d_k, y_j), \ \forall \ d_k \in \vec{P_{s_j i}} \tag{6.11}$$

As illustrated, $Rep(P_{s_ji}, y_i)$ denotes the reputation score of the peer P_i excluding the interaction data of testing query topic y_i. The robustness of scenario 2 resides in excluding the interaction data of the training queries generated by the testing query..

Based on the two discussed scenarios the research question is:

  - **RQ-6.1:** Could the effectiveness of P2P-IR systems be improved by using reputation measures mined from past interaction data?

- **Scenario 3 (or (Training-Testing)%)**: leave-one-out method effectively excludes interaction data associated with testing topic. However, in order to evaluate the robustness of the reputation-based methods on the amount of training data, I developed scenario 3. In the simulation, I have the information of interaction data for each query topic $y_i \in Y$ separately. Through these interaction data, scenario 3 stress testing the system and estimates the reputation score of the peers by excluding more interaction data. In other words, the super-peers exclude a various proportion of training queries and their interaction data as discussed in scenario 2. The idea is to understand the behaviour of the system with limited interaction data. The percentages used in this dissertation are 25%, 50%, and 75%. As an example, I estimate the reputation score of a peer P_{s_ji} by excluding just $Q^{y_i}_{training}$ interaction data from 25% of query topics in Y. I use the remaining 75% of the interaction data for the testing phase. This leave-out method is called (25-75)% as stated with notation (Training-Testing)%. The other leave-out methods used in this scenario are (50-50)% and (75-25)%. The research question in this scenario is:

  - **RQ-6.2:** How effective is the retrieval effectiveness with varying amount of interaction data?

- **Scenario 4 (or Noisy interaction data)**: In the previous scenarios, I assumed that the users click and download only the relevant documents. But what about if the user clicked or downloaded non-relevant documents in the final merged ranked list. In scenario 4, the evaluation depends on the interaction data of the users not just on relevant documents but also on

non-relevant documents in the final merged result list. Specifically, in the simulation of user interaction, I select up to five percentages of random non-relevant documents (or noisy data) from the merged ranked list as being clicked or downloaded by a user. The users on the web have significantly little satisfaction on clicking non-relevant documents using the queries they rated on their past interactions (Carterette and Jones, 2007; Xing et al., 2013). Based on this assumption, I use these percentage values, which are 1%, 2%, 3%, 4%, and 5% of non-relevant documents of the merged result list assuming that the users click and/or download non-relevant documents in a low probability. The feedback values of the non-relevant documents are randomly assigned between 1 and 4. The research question using noisy interaction data is:

- **RQ-6.3:** How does the Reputation-based approach perform under noisy data?

However, I study these fourth scenarios in the proposed approaches except the RCORI and RPCORI methods. I did separate experiments to study the retrieval effectiveness of using CORI approach instead of simple term frequency statistical information. In this case, I investigate the ability to apply the CORI approach to improving the effectiveness or not. The research question is:

- **RQ-6.4:** Is there any effect of using the CORI resource selection method as a content-based statistic information to rank and select the most reputable peers?

6.3.2 Retrieval Efficiency

The retrieval efficiency depends on a number of factors: the delay of messages in the network; to the response time of the retrieval models at each peer; to the time retrieving and merging the result for a given query. In this chapter, I discuss the efficiency from the message complexity perspective, which is the number of messages sent across the P2P network. The benefits of using message complexity criterion are to correlate the retrieval effectiveness with the required amount of messages in the system and examine how robust are the reputation-based approaches compared with the baselines approaches. For evaluation, I use

five percentages of selected peers at each super-peer, which are 10%, 20%, 30%, 40%, and 50%. As the number of peers at each super-peer vary, I expect a different number of messages sent for different retrieval approaches. Given the message complexity as efficiency criterion, the research question is:

- **RQ-6.5:** Does the reputation-based P2P-IR approach improve the retrieval efficiency in terms of network traffic?

6.4 Experimental Results

The proposed methods are assessed based on retrieval effectiveness and efficiency measures. In the retrieval effectiveness measures, I use the following IR evaluation metrics: Precision; Recall; P@10; P@30; P@100; and MAP. The efficiency measures (or message complexity) are based on the number of messages routed to the selected peers as I will discuss later.

6.4.1 Retrieval effectiveness

For the retrieval effectiveness perspective, the statistically significant improvements are measured using the hypothetical two-paired bootstrap t-test, which is the most powerful of significant tests in IR retrieval scenarios (Urbano et al., 2013). As I have two baselines, which are CORI and Taily approaches, the statistically significant improvements at $p \leq 0.05$ denoted as $\uparrow$; while significant improvements at $p \leq 0.01$ are denoted $\Uparrow$ in comparison with Taily selection method. Similarly, the statistically significant degradations are denoted $\downarrow$ and $\Downarrow$, respectively. The statistically significant improvements in comparison with the CORI selection method at $p \leq 0.05$ denoted as $\rightarrow$; while significant improvements at $p \leq 0.01$ are denoted $\Rightarrow$. Similarly, the statistically significant degradations are denoted $\leftarrow$ and $\Leftarrow$, respectively. The comparison between R and RP methods and RT and RPT methods are conducted under statistical significant measures using the same hypothetical two-paired bootstrap t-test, where the statistically significant improvements at $p \leq 0.05$ denoted as $\triangle$; while significant improvements at $p \leq 0.01$ are denoted $\blacktriangle$. Similarly, the statistically significant degradations are denoted $\triangledown$ and $\blacktriangledown$, respectively. The best retrieval value for specific measurement is highlighted in boldface.

Table 6.1: Reputation-based effectiveness: Scenario 1 at 10% of Selected Peers

DL*	**DLWOR Testbed**						**DLWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
CORI	0.02588	0.47515	0.179	0.13167	0.0871	0.095	0.0127	0.1242	0.032	0.02967	0.0256	0.00978
Taily	0.02769	0.50463	0.182	0.137	0.0898	0.09493	0.02523	0.46911	0.026	0.025	0.0331	0.02978
R	0.02532$^{\downarrow}$	0.42297$^{\leftarrow\Downarrow}$	0.137$^{\leftarrow\Downarrow}$	0.128	0.0819$^{\Downarrow}$	0.06987$^{\Leftarrow\Downarrow}$	**0.02228**$^{\Rightarrow\Downarrow}$	**0.3711**$^{\Rightarrow\Downarrow}$	**0.027**	**0.02633**	**0.0296**	**0.02307**$^{\Rightarrow\Downarrow}$
RT	0.02532$^{\downarrow}$	0.42297$^{\leftarrow\Downarrow}$	0.137$^{\leftarrow\Downarrow}$	0.128	0.0819$^{\Downarrow}$	0.06987$^{\Leftarrow\Downarrow}$	**0.02228**$^{\Rightarrow\Downarrow}$	**0.3711**$^{\Rightarrow\Downarrow}$	**0.027**	**0.02633**	**0.0296**	**0.02307**$^{\Rightarrow\Downarrow}$
RP	**0.0346**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.551**$^{\Rightarrow\blacktriangle}$	**0.23**$^{\blacktriangle}$	**0.22967**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1572**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15766**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02631**$^{\Rightarrow\blacktriangle}$	**0.43545**$^{\Rightarrow\blacktriangle}$	**0.139**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.125**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0904**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07507**$^{\Rightarrow\Uparrow\blacktriangle}$
RPT	**0.0346**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.551**$^{\Rightarrow\blacktriangle}$	**0.23**$^{\blacktriangle}$	**0.22967**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1572**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15766**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02631**$^{\Rightarrow\blacktriangle}$	**0.43545**$^{\Rightarrow\blacktriangle}$	**0.139**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.125**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0904**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07507**$^{\Rightarrow\Uparrow\blacktriangle}$
ASIS*	**ASISWOR Testbed**						**ASISWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
CORI	0.0265	0.4767	0.164	0.122	0.0801	0.07699	0.0233	0.44764	0.015	0.02	0.023	0.02272
Taily	0.02613	0.45867	0.159	0.12233	0.0785	0.07103	0.02088	0.39249	0.02	0.02133	0.0233	0.02218
R	**0.02816**$^{\uparrow}$	**0.45929**	**0.175**$^{\Uparrow}$	**0.13033**$^{\uparrow}$	**0.0828**$^{\Uparrow}$	**0.07745**$^{\uparrow}$	**0.02213**	**0.3948**$^{\Leftarrow}$	**0.022**$^{\Rightarrow}$	**0.027**$^{\rightarrow\uparrow}$	**0.02430**	0.02203
RT	**0.02816**$^{\uparrow}$	**0.45929**	**0.175**$^{\Uparrow}$	**0.13033**$^{\uparrow}$	**0.0828**$^{\Uparrow}$	**0.07745**$^{\uparrow}$	**0.02213**	**0.3948**$^{\Leftarrow}$	**0.022**$^{\Rightarrow}$	**0.027**$^{\rightarrow\uparrow}$	**0.0243**	0.02203
RP	**0.0418**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.669**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.236**$^{\rightarrow\Uparrow\vartriangle}$	**0.23767**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1782**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17901**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.028**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49608**$^{\rightarrow\Uparrow\blacktriangle}$	**0.108**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09234**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0687**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06187**$^{\Rightarrow\Uparrow\blacktriangle}$
RPT	**0.0418**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.669**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.236**$^{\rightarrow\Uparrow\vartriangle}$	**0.23767**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1782**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17901**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.028**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49608**$^{\rightarrow\Uparrow\blacktriangle}$	**0.108**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09234**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0687**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06187**$^{\Rightarrow\Uparrow\blacktriangle}$
U*	**UWOR Testbed**						**UWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
CORI	0.02841	0.50793	0.19	0.14667	0.0955	0.10819	0.02808	0.49536	0.015	0.02433	0.0363	0.03196
Taily	0.02781	0.48616	0.183	0.14633	0.0932	0.10931	0.02563	0.45538	0.014	0.02067	0.0292	0.02751
R	**0.03474**$^{\uparrow}$	0.40175$^{\Leftarrow\Downarrow}$	**0.285**$^{\Rightarrow\Uparrow}$	**0.19766**$^{\Rightarrow\Uparrow}$	**0.1125**$^{\Rightarrow\Uparrow}$	**0.12524**$^{\rightarrow\Uparrow}$	**0.03008**	0.36649$^{\Leftarrow\Downarrow}$	**0.031**$^{\Rightarrow\Uparrow}$	**0.03633**$^{\rightarrow\uparrow}$	**0.0334**	0.0259$^{\leftarrow}$
RT	**0.03478**$^{\uparrow}$	0.40357$^{\Leftarrow\Downarrow}$	**0.289**$^{\Rightarrow\Uparrow}$	**0.199**$^{\Rightarrow\Uparrow}$	**0.1129**$^{\Rightarrow\Uparrow}$	**0.12713**$^{\Rightarrow\Uparrow}$	**0.03009**	0.36694$^{\Leftarrow\Downarrow}$	**0.031**$^{\Rightarrow\Uparrow}$	**0.03633**$^{\rightarrow\uparrow}$	**0.0334**	0.02594$^{\leftarrow}$
RP	**0.03939**$^{\Rightarrow\Uparrow\blacktriangle}$	0.47375$^{\blacktriangle}$	**0.247**$^{\uparrow}$	**0.22533**$^{\Rightarrow\Uparrow}$	**0.1518**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13164**$^{\Uparrow}$	**0.0325**$^{\Uparrow\blacktriangle}$	0.40733$^{\Leftarrow\blacktriangle}$	**0.121**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11033**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0814**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06268**$^{\Rightarrow\Uparrow\blacktriangle}$
RPT	**0.03943**$^{\Rightarrow\Uparrow\blacktriangle}$	0.47557$^{\blacktriangle}$	**0.247**$^{\uparrow}$	**0.22633**$^{\Rightarrow\Uparrow}$	**0.1522**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13265**$^{\Uparrow}$	**0.03251**$^{\Rightarrow\Uparrow\blacktriangle}$	0.40779$^{\Leftarrow\blacktriangle}$	**0.121**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11033**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0814**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06272**$^{\Rightarrow\Uparrow\blacktriangle}$

As discussed in scenario 1, the reputation scores of the peers at current testing query are estimated over all interaction data of training queries that are *generated* by all testing query topics. Table 6.1 displays the retrieval effectiveness results in scenario 1 on three environments using 10% of the selected peers at each super-peer using the proposed approaches, which are R, RT, RP, and RPT methods. RP and RPT methods significantly outperform the baselines and their corresponding R and RT methods in almost all the cases over the testbeds of the three environments. In DL^* testbeds, R and RT methods show comparable and better performance in some cases, especially in DLWR testbed compared with the CORI method. Across tesbeds , the results of reputation-based methods show high effective values in DLWR testbed compared to the DLWOR testbed, due to the replication of relevant documents in the testbed. This replication strategy affects the reputation score of the peers as the relevant documents have more chance to appear in the final ranked list with more probability of downloading or clicking by the users. The results in $ASIS^*$ testbeds are totally different, the R and RT methods obtain significant retrieval performance in most of the cases compared to the Taily method and competitive results to the CORI approach. The uniformly distributed environment shows significant results for R and RT methods in some metrics with comparable results on most of the cases compared

to the two baseline methods.

A further selection of peers improves the retrieval effectiveness of the proposed approaches as shown in Tables B.1, B.2, and B.3 for the other percentage of values. In DL^* family as shown in Table B.1, still RP and RPT methods gain significant improvements in all the cases of measurement metrics, while R and RT methods have comparable results in a few cases in DLWOR and more in DLWR at 20% and 30% of selected peers. The better improvements in DLWR testbed occurs due to replication strategy along further with much better performance at 40% and 50% on both testbeds. Table B.2 shows significant results for the proposed methods in approximately all the cases under the file sharing environments (i.e, $ASIS^*$ family). This gives an indication of the robustness of the reputation-based approaches in selecting the relevant peers from a large number of peers. The uniformly distributed environment in Table B.3 behaves much closer in retrieval performance to the ASIS environment with significant retrieval results to nearly 99% of the measurement metrics. Results in scenario 1 show a comparable or significant results for the proposed reputation-based methods and show their ability in environments of a large number of peers (i.e, ASIS and U environments) to improve the retrieval effectiveness. The question is, *does including the training queries' interaction data from the testing query topic give significant impact on retrieval effectiveness using the reputation-based approaches in scenario 1?*

Table 6.2 shows the retrieval effectiveness results using scenario 2 that excludes the training queries' interaction data from the testing query topic for the estimation of peers' reputation values, which is the leave-one-out method. In DL^* family, the proposed approaches obtain competitive or significantly better retrieval effectiveness in almost all the cases, especially in the RP and RPT methods. In particular, specifically for non-prioritizing approaches, the R and RT methods show significant results in DLWR to Precision, Recall, and MAP metrics in comparison with the CORI method and competitive results to the other metrics on the same testbed for both the baseline methods. In more details with a number of selected peers as shown in Table B.4, R method , in DLWOR testbed, achieves high retrieval quality to Precision and P@100 metrics at 40% and significant value to Recall and P@100 (in comparison with the CORI approach) at 50% of selected peers. In addition, R method gains better retrieval

Table 6.2: Reputation-based effectiveness: Scenario 2 at 10% of Selected Peers

DL*	**DLWOR Testbed**						**DLWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
CORI	0.02588	0.47515	0.179	0.13167	0.0871	0.095	0.0127	0.1242	0.032	0.02967	0.0256	0.00978
Taily	0.02769	0.50463	0.182	0.137	0.0898	0.09493	0.02523	0.46911	0.026	0.025	0.0331	0.02978
R	0.02483$^{\downarrow}$	0.4166$^{\leftarrow\Downarrow}$	0.122$^{\Leftarrow\Downarrow}$	0.115$^{\Downarrow}$	0.077$^{\leftarrow\Downarrow}$	0.0643$^{\Leftarrow\Downarrow}$	**0.02229**$^{\Rightarrow\Downarrow}$	**0.36744**$^{\Rightarrow\Downarrow}$	**0.027**	**0.026**	**0.0284**	**0.02222**$^{\Rightarrow\Downarrow}$
RT	0.02612	0.44371$^{\Downarrow}$	0.122$^{\Leftarrow\Downarrow}$	0.11866$^{\Downarrow}$	0.0804$^{\Downarrow}$	0.06805$^{\Leftarrow\Downarrow}$	**0.02415**$^{\Rightarrow}$	**0.41621**$^{\Rightarrow\downarrow}$	**0.028**	**0.028**	**0.0312**	**0.026**$^{\Rightarrow\downarrow}$
RP	**0.03375**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.54**$^{\Rightarrow\blacktriangle}$	**0.215**$^{\blacktriangle}$	**0.216**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1494**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14496**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02556**$^{\Rightarrow\blacktriangle}$	**0.41835**$^{\Rightarrow\downarrow\blacktriangle}$	**0.125**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11166**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0821**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06431**$^{\Rightarrow\Uparrow\blacktriangle}$
RPT	**0.03786**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.597**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.276**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.26933**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1829**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1998**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03191**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.52681**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.186**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18633**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1362**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13176**$^{\Rightarrow\Uparrow\blacktriangle}$
ASIS*	**ASISWOR Testbed**						**ASISWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
CORI	0.02650	0.4767	0.164	0.122	0.0801	0.077	0.0233	0.44764	0.015	0.02	0.023	0.02272
Taily	0.02613	0.45867	0.159	0.12233	0.0785	0.07103	0.02088	0.39249	0.02	0.02133	0.0233	0.02218
R	**0.02805**$^{\uparrow}$	0.4578	**0.175**$^{\Uparrow}$	**0.13**$^{\uparrow}$	**0.0826**$^{\Uparrow}$	**0.07702**$^{\uparrow}$	**0.02208**	**0.39322**$^{\Leftarrow}$	**0.022**$^{\Rightarrow}$	**0.027**$^{\rightarrow\uparrow}$	**0.0243**	**0.02202**
RT	**0.02842**$^{\rightarrow\Uparrow}$	**0.47268**	**0.175**$^{\Uparrow}$	**0.13**$^{\uparrow}$	**0.0829**$^{\Uparrow}$	**0.07797**$^{\uparrow}$	**0.02226**	**0.39705**$^{\Leftarrow}$	**0.022**$^{\Rightarrow}$	**0.027**$^{\rightarrow\uparrow}$	**0.0243**	**0.02238**
RP	**0.04093**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.224**$^{\uparrow\rightarrow}$	**0.223**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1668**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16345**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02719**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.48247**$^{\Uparrow}$	**0.103**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08367**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0621**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05272**$^{\Rightarrow\Uparrow\blacktriangle}$
RPT	**0.04475**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.725**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.272**$^{\Rightarrow\Uparrow\triangle}$	**0.265**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1994**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.2202**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03271**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56046**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.138**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13166**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1078**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10027**$^{\Rightarrow\Uparrow\blacktriangle}$
U*	**UWOR Testbed**						**UWR Testbed**					
Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
CORI	0.02841	0.50793	0.19	0.14667	0.0955	0.10819	0.02808	0.49536	0.015	0.02433	0.03630	0.03196
Taily	0.02781	0.48616	0.183	0.14633	0.0932	0.10931	0.02563	0.45538	0.014	0.02067	0.0292	0.02751
R	**0.03292**	0.37312$^{\Leftarrow\Downarrow}$	**0.266**$^{\Rightarrow\Uparrow}$	**0.18066**$^{\Rightarrow\Uparrow}$	**0.1036**	**0.11027**	**0.02892**	0.34463$^{\Leftarrow\Downarrow}$	**0.029**$^{\Rightarrow\Uparrow}$	**0.035**$^{\uparrow}$	**0.0316**	0.02372$^{\leftarrow}$
RT	**0.03297**	0.37539$^{\Leftarrow\Downarrow}$	**0.271**$^{\Rightarrow\Uparrow}$	**0.18233**$^{\Rightarrow\Uparrow}$	**0.1041**	**0.11258**	**0.02895**	0.34600$^{\Leftarrow\Downarrow}$	**0.029**$^{\Rightarrow\Uparrow}$	**0.035**$^{\uparrow}$	**0.0316**	0.02385$^{\leftarrow}$
RP	**0.03728**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.43933**$^{\rightarrow\blacktriangle}$	**0.228**	**0.20233**$^{\Rightarrow\Uparrow}$	**0.136**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11165**	**0.03104**$^{\uparrow\blacktriangle}$	**0.37958**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.114**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09767**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0728**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05288**$^{\Rightarrow\Uparrow\blacktriangle}$
RPT	**0.03733**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.4416**$^{\rightarrow\blacktriangle}$	**0.231**	**0.20367**$^{\Rightarrow\Uparrow}$	**0.1364**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11274**	**0.03107**$^{\uparrow\blacktriangle}$	**0.38095**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.114**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09767**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0728**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05297**$^{\Rightarrow\Uparrow\blacktriangle}$

result to Recall metric and significant results to P@100 (comparing to the two baselines) and MAP (comparing to the Taily method) metrics. RT method has high retrieval effectiveness to Precision at 30% and Recall at 40% with significantly better values to Precision at 40% (comparing to the CORI approach) and 50% (in comparison with the Taily method). In DLWR testbed, both R and RT methods gain significant improvements in almost all the cases compared to the CORI approach as well as better values in comparison with the Taily approach. Through using content information for routing, the RT and RPT methods exhibit significant results on both testbeds over almost all metrics. In comparing the prioritizing approaches RP and RPT methods with their corresponding approaches; R and RT, we can see significant improvement values in all cases of the testbeds due to the promotion of relevant and reputable documents at the top of the final merged results.

In file-sharing environments represented by $ASIS^*$ family, the retrieval results are much better than the DL environment. As shown, the R and RT methods obtain significant improvements in approximately almost all the cases and with competitive values in other metrics. As shown also, RT approach outperforms the R method with higher values in all the cases on both testbeds at 10% of the selected peers, whereas they have the same stable values of the cases between

20% and 50% of the selected peers as shown in Table B.5. RP and RPT methods increase the retrieval effectiveness significantly in comparison with the two baselines and their corresponding methods (R and RP). Additionally, both RP and RPT methods have stable values between 20% to 50% of the selected peers. This is due to the fact that the reputation-based methods incorporate usage information, which follows a power-law pattern seen in real-life file-sharing scenarios and depicted by this ASIS* family testbeds. Since this pattern appears to give a stable performance to the proposed approaches as the number of reputable peers are the same in all the percentages.

The proposed methods achieve significant improvements in almost all the cases of uniformly distributed environments. RP and RPT methods have high improvement to P@10 metric on all percentages and better values to MAP metric at 10% of selected peers. RP and RPT approaches are not significantly higher in values to R and RT methods on P@10 and MAP metrics. Moreover, we can see metrics stabilise values between 40% and 50% of selected peers as shown in Table B.6.

In summary, however, R and RT methods obtain significantly better improvement for $ASIS^*$ and U^* families. This topical (relevant) documents are distributed differently to particular peers in $ASIS^*$ and U^* families. While these topics are condensed on a small number of peers in DL^* environment making an effect on selection performance of this family. The RP and RPT methods gain highly significant improvement in almost all testbeds over the three environments. Such results obtained due to promoting those relevant and/or reputable documents at the top of the final merged result. In both scenarios 1 and 2, the reputation-based approaches give better and significant results in almost all the cases and answer the research question **RQ-6.1** of retrieval effectiveness of the proposed methods as query routing in semi-structured P2P-IR networks.

6.4.2 Effectiveness of Varying Training and Testing Boundaries

I used the 1-leave-out method in the evaluation of proposed methods (i.e, R, RT, RP, RPT) in scenarios 2. When the testing query arrives, a super-peer rules out the training usage information that is associated with the testing query topic.

This subsection studies the other three leave-out methods; which are (25-75)%, (50-50)%, and (75-25)% as discussed in scenario 3. In a nutshell, the first number of each method is the percentage of training usage information to leave from the calculation regardless of the topics they have. The second number refers to the rest of the training queries that will be used for the testing phase. Therefore, the query of leave-out training information are not used in the calculation and the results are averaged over solely on testing queries. I have a lot of results with three leave-out methods on six testbeds with five percentages of selected peers between 10% to 50% (i.e, 90 tables), hence I summarise only the average results over percentage of selected peers.

Table 6.3: Retrieval effectiveness DL Environment 10% of Selected Peers

	DL*	DLWOR Testbed						DLWR Testbed					
(Train-Test)%	Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
(25-75)%	Taily	0.02805	0.51313	0.187	0.13644	0.09307	0.09359	0.02516	0.47292	0.013	0.01689	0.02773	0.02624
	CORI	0.02590	0.46086	0.18	0.13156	0.08987	0.09323	0.01123	0.11786	0.024	0.02578	0.02413	0.00943
	R/RP	0.02085	0.35184	0.172	0.12044	0.07493	0.06135	**0.01933**$^{\Rightarrow}$	**0.32426**$^{\Rightarrow}$	**0.017**$^{\Uparrow}$	**0.01777**$^{\uparrow}$	0.02240	**0.01795**$^{\Rightarrow}$
	RT	**0.03333**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.59057**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.223**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.176**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11427**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1185**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02955**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.54849**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.021**$^{\Uparrow\blacktriangle}$	**0.02355**$^{\Uparrow\blacktriangle}$	**0.0348**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03638**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04314**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.71657**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.475**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.39111**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.232**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.35209**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03615**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.63976**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.267**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.236**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15413**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17583**$^{\Rightarrow\Uparrow\blacktriangle}$
(50-50)%	Taily	0.02867	0.48933	0.184	0.14066	0.0928	0.08805	0.02524	0.44015	0.01	0.01066	0.0218	0.02142
	CORI	0.02583	0.42826	0.184	0.13667	0.0886	0.08318	0.00946	0.10416	0.014	0.02	0.02480	0.00652
	R/RP	0.02115	0.33267	0.176	0.126	0.0784	0.06334	0.02002	**0.31789**$^{\Rightarrow}$	**0.012**	**0.01399**$^{\uparrow}$	0.0164	**0.01576**$^{\Rightarrow}$
	RT	**0.03435**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55745**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.228**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17933**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1154**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11451**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03061**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49930**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.016**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.01533**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0252**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03104**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04489**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.6921**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.502**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.41867**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.252**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34932**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03883**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.60584**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.282**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25267**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1688**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17568**$^{\Rightarrow\Uparrow\blacktriangle}$
(75-25)%	Taily	0.03499	0.4924	0.184	0.164	0.11	0.1115	0.03019	0.42772	0.016	0.01733	0.026	0.02661
	CORI	0.03185	0.42236	0.192	0.15333	0.1056	0.09588	0.01235	0.10771	0.02	0.01867	0.0244	0.00807
	R/RP	0.02573	0.33246	**0.2**$^{\uparrow}$	**0.15466**	0.092	0.08621	**0.02516**$^{\Rightarrow}$	**0.32562**$^{\Rightarrow}$	**0.02**$^{\uparrow}$	**0.02**	0.01720	**0.02127**$^{\Rightarrow}$
	RT	**0.04073**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.57**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24**$^{\Rightarrow\Uparrow\triangle}$	**0.196**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1264**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.036**$^{\Rightarrow\uparrow\blacktriangle}$	**0.50333**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.024**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02**$^{\Rightarrow\uparrow\blacktriangle}$	**0.0244**$^{\blacktriangle}$	**0.0355**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05113**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.69452**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.456**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.38667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.2364**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.35872**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.04451**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.61181**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.284**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24134**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.166**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19847**$^{\Rightarrow\Uparrow\blacktriangle}$

Table 6.3 shows the experimental results of the leave-out methods on the DL environment on 10% of selected peers. Overall, the proposed approaches show significant retrieval effectiveness improvements over the baseline methods in almost all the cases and much better performance for further percentages of peer selection as shown in Tables B.7, B.8, and B.9. RT and RPT methods consistently obtain better retrieval effectiveness results to the baseline methods and the other two proposed R and RP approaches (except RT method at P@10 on DLWR testbed comparing to CORI approach for (25-75)% leave-out method). Specifically in DLWR testbed, R and RP methods obtain high and competitive values over all leave-out approaches in comparison with the CORI algorithm, whereas RT and RPT methods obtain better performance over the two baseline methods. In spite of poor usage information of documents' reputation values, the reputation-based approaches that depend on the content-based information

in their calculation (i.e, RT and RPT methods) can route a query to the relevant peers. The other two reputation-based approaches (i.e, R and RP methods) have an effect on retrieval effectiveness using these leave-out methods, but they still obtain better and competitive results in some cases especially in DLWR testbed.

In $ASIS^*$ testbeds as shown in Table 6.4 and the other percentages of selected peers shown in Tables B.10, B.11, and B.12, the proposed approaches show better retrieval performance in almost all measurement metrics in comparison with the CORI and the Taily approaches, although the testbeds in such family have a small number of documents on average for each peer. This reduction in a number of documents could have an effect on past usage information when using the training data information of the given testing query. But as seen, even with poor usage information, the reputation-based methods perform well better than the baseline methods.

Table 6.4: Retrieval effectiveness ASIS Environment 10% of Selected Peers

	ASIS*	**ASISWOR Testbed**						**ASISWR Testbed**					
(Train-Test)%	Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
(25-75)%	Taily	0.0262	0.45595	0.159	0.12533	0.08293	0.06319	0.01989	0.39129	0.009	0.01467	0.01907	0.01849
	CORI	0.02666	0.46539	0.157	0.12044	0.08307	0.06748	0.02275	0.43106	0.009	0.01378	0.01853	0.01975
	R/RP	**0.0276**$^{\uparrow}$	0.44221	**0.173**$^{\Rightarrow\Uparrow}$	**0.13378**$^{\Rightarrow\Uparrow}$	**0.08693**$^{\Rightarrow\Uparrow}$	**0.06747**$^{\Uparrow}$	**0.02074**$^{\uparrow}$	0.37376	**0.016**$^{\Rightarrow\Uparrow}$	**0.02**$^{\Rightarrow\Uparrow}$	**0.0188**	**0.01882**
	RT	**0.02893**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.47291**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.173**$^{\Rightarrow\Uparrow}$	**0.13511**$^{\Rightarrow\Uparrow}$	**0.0888**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07102**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02089**$^{\Rightarrow\Uparrow}$	0.37979	**0.016**$^{\Rightarrow\Uparrow}$	**0.02**$^{\Rightarrow\Uparrow}$	**0.01907**$^{\Rightarrow\blacktriangle}$	**0.01924**$^{\Uparrow\blacktriangle}$
	RPT	**0.04117**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.66399**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.443**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.36623**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.21307**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29164**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02977**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.51761**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.131**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12889**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.096**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09388**$^{\Rightarrow\Uparrow\blacktriangle}$
(50-50)%	Taily	0.0272	0.45007	0.154	0.12667	0.0852	0.05847	0.02046	0.3643	0.002	0.008	0.0128	0.0131
	CORI	0.0276	0.45947	0.156	0.122	0.0858	0.06241	0.02366	0.41009	0.002	0.008	0.01340	0.01513
	R/RP	**0.02866**$^{\Rightarrow\Uparrow}$	0.4468	**0.17**$^{\Rightarrow\Uparrow}$	**0.136**$^{\Rightarrow\Uparrow}$	**0.0894**$^{\Rightarrow\Uparrow}$	**0.06421**$^{\rightarrow\Uparrow}$	0.02042	0.34322	**0.006**$^{\Rightarrow\Uparrow}$	**0.012**$^{\Rightarrow\Uparrow}$	**0.013**$^{\Uparrow}$	0.01231
	RT	**0.02985**$^{\Rightarrow\Uparrow\vartriangle}$	**0.47363**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17**$^{\Rightarrow\Uparrow}$	**0.13667**$^{\Rightarrow\Uparrow}$	**0.0912**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06725**$^{\Rightarrow\Uparrow\vartriangle}$	**0.02066**$^{\Rightarrow\Uparrow}$	0.35263	**0.006**$^{\Rightarrow\Uparrow}$	**0.012**$^{\Rightarrow\Uparrow}$	**0.013**$^{\Uparrow}$	0.01274
	RPT	**0.04331**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.66163**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.472**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.40067**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.2362**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.2985**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.0312**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50807**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.118**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13334**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.1046**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08431**$^{\Rightarrow\Uparrow\blacktriangle}$
(75-25)%	Taily	0.03348	0.44831	0.188	0.156	0.1004	0.06998	0.02303	0.34826	0.004	0.012	0.01520	0.01492
	CORI	0.03416	0.47295	0.192	0.14667	0.1004	0.07852	0.02775	0.40698	0.004	0.012	0.0176	0.01809
	R/RP	**0.03608**$^{\Rightarrow\Uparrow}$	0.4429	**0.188**	**0.16**$^{\Rightarrow\Uparrow}$	**0.104**	**0.07462**	**0.02325**	0.32739	**0.012**$^{\Rightarrow\Uparrow}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.0164**$^{\Uparrow}$	0.01404
	RT	**0.03767**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.47936**$^{\Uparrow\blacktriangle}$	**0.188**	**0.16**$^{\Rightarrow\Uparrow}$	**0.106**	**0.0782**$^{\Uparrow}$	**0.02331**	0.33344	**0.012**$^{\Rightarrow\Uparrow}$	**0.017**$^{\Rightarrow\Uparrow}$	**0.0164**$^{\Uparrow}$	0.01453
	RPT	**0.051**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.66857**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.428**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.37467**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.2268**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29896**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03471**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49608**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.096**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12534**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.094**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09241**$^{\Rightarrow\Uparrow\blacktriangle}$

In Uniformly distributed environments in Tables 6.5, B.13, B.14, and B.15, show retrieval performance for the proposed methods on all leave-out approaches with almost stable values in some of the cases compared to the two baseline methods.

In summary, the results give evidence that the reputation-based resource selection methods, even with small amount of past usage data information, can route the given query effectively to the relevant peers. In addition, the results show highly effective results to the proposed approaches in ASIS and U environments due to the relevant document distribution on different peers, which does not have any effect on retrieval effectiveness of excluding the various amounts

Table 6.5: Retrieval effectiveness U Environment 10% of Selected Peers

	U*	UWOR Testbed						UWR Testbed					
(Train-Test)%	Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
(25-75)%	Taily	0.02774	0.5	0.185	0.152	0.09693	0.10361	0.02554	0.46802	0.008	0.01644	0.02387	0.02477
	CORI	0.02828	0.50584	0.192	0.152	0.0992	0.10065	0.02804	0.49391	0.011	0.02177	0.03187	0.03006
	R/RP/RT/RPT	0.02758	0.35843	**0.264**$^{\Rightarrow\Uparrow}$	**0.173**$^{\Rightarrow\Uparrow}$	**0.10053**	**0.10144**	0.02487	0.34536	**0.024**$^{\Rightarrow\Uparrow}$	**0.024**$^{\Rightarrow\Uparrow}$	**0.02387**	0.01956
(50-50)%	Taily	0.02765	0.4599	0.192	0.16267	0.1006	0.10115	0.02639	0.43953	0.002	0.014	0.01620	0.02119
	CORI	0.02853	0.47053	0.196	0.16134	0.102	0.09531	0.02905	0.46653	0.006	0.01866	0.0298	0.02838
	R/RP/RT/RPT	0.02585	0.35044	**0.276**$^{\Rightarrow\Uparrow}$	**0.184**$^{\Rightarrow\Uparrow}$	**0.1118**	**0.10141**$^{\Rightarrow}$	0.02409	0.334	**0.014**$^{\Rightarrow\Uparrow}$	**0.01933**$^{\Rightarrow\Uparrow}$	**0.022**$^{\Uparrow}$	0.01696
(75-25)%	Taily	0.03223	0.43508	0.224	0.17733	0.11	0.12231	0.03143	0.433	0.004	0.02133	0.018	0.02383
	CORI	0.03471	0.45954	0.24	0.18134	0.118	0.11394	0.03559	0.46443	0.004	0.02266	0.024	0.03098
	R/RP/RT/RPT	**0.03321**$^{\uparrow}$	0.39039	**0.276**$^{\Rightarrow\Uparrow}$	**0.20534**$^{\Rightarrow\Uparrow}$	**0.1304**$^{\Rightarrow\Uparrow}$	**0.11993**	0.02964	0.3355	**0.016**$^{\Rightarrow\Uparrow}$	**0.02133**	**0.0188**	0.01764

of usable information. This answered the research question (**RQ-6.2**) as with varying training data we can get retrieval quality improvements.

6.4.3 Reputation-based Approaches Under Noisy Data

Users of IR systems sometimes, in an indiscernible manner, click and download documents that do not cater their information need. The reason is that users expect from the IR searching systems to retrieve high-quality documents. But the IR systems use probabilistic retrieval models to assign scores to the documents on a different indication of relevance between models. Hence, some relevant documents might match the query terms and own a score in the retrieved result list. Perusing this list, the users may click or download non-relevant documents. In addition, a document retrieved on a specific query might be relevant to user A and non-relevant to user B due to the subjective nature of the information need of such query. Therefore, I conduct an experiment to simulate the behaviour of users when they click or download noisy data (or non-relevant documents).

Table 6.6 shows retrieval effectiveness results based on the scenario 4 including noisy past interaction information. The table presents the results at 1% of noisy information of the reputation-based approaches (where the method name in the table is appended with _N). The baseline approaches in this scenario are the corresponding reputation-based approaches that discussed in scenario 2 as shown in baseline row. I did not use significant test here as the results of reputation methods with noisy data compare to their corresponding methods do not obtain much differences. In particular, the RP_N approach outperforms its corresponding baseline approaches (i.e, RP) in all the measurement metrics on the three environments (except with the competitive result at P@10 for DLWOR

Table 6.6: Retrieval Effectiveness on 1% of Noisy Data

DL*		DLWOR testbed						DLWR testbed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
Baseline	R	0.02822	0.50753	0.1182	0.1212	0.08628	0.07478	0.02468	0.46074	**0.0262**	0.0234	**0.02612**	**0.02512**
	RP	0.04062	0.69047	**0.2242**	0.23126	0.17394	0.18629	0.03029	0.54793	0.1156	0.1068	0.08458	0.07452
	RT	**0.02885**	**0.52324**	0.119	0.12313	**0.08764**	0.07672	**0.02566**	**0.48592**	0.0256	0.02386	0.02732	**0.02708**
	RPT	**0.04479**	**0.75092**	**0.2836**	**0.28526**	**0.20832**	**0.24972**	**0.03689**	**0.65343**	**0.1702**	**0.1746**	**0.14092**	**0.14557**
1%	R_N	0.02822	**0.50903**	0.1182	**0.1216**	**0.08634**	**0.07498**	0.02463	**0.46085**	0.02520	0.02340	0.02568	0.0249
	RP_N	**0.13658**	**0.72589**	0.224	**0.23306**	**0.17792**	**0.19772**	**0.13551**	**0.57992**	**0.1702**	**0.1746**	**0.14272**	**0.13851**
	RT_N	0.02613	0.44726	**0.1492**	**0.12393**	0.0806	**0.07826**	0.02381	0.41334	**0.0264**	**0.02513**	**0.0275**	0.02437
	RPT_N	0.01827	0.25159	0.0688	0.07066	0.05126	0.03283	0.01945	0.2712	0.0276	0.0298	0.02878	0.02033
ASIS*		ASISWOR testbed						ASISWR testbed					
Baseline	R	**0.02887**	0.47723	**0.17100**	**0.1292**	**0.08388**	**0.07822**	**0.02216**	**0.39499**	0.022	0.02673	0.02402	**0.02212**
	RP	0.04221	0.68605	**0.2248**	0.22407	0.16848	0.1681	0.02741	0.48757	0.1006	0.08313	0.06092	0.05249
	RT	**0.02895**	**0.48021**	**0.171**	**0.1292**	**0.08394**	**0.07841**	0.0222	**0.39576**	0.022	0.02673	**0.02402**	0.02219
	RPT	**0.04592**	**0.74608**	**0.2728**	**0.26553**	**0.20068**	**0.2243**	**0.0328**	**0.56216**	**0.1364**	**0.1302**	**0.1065**	**0.09922**
1%	R_N	0.02886	**0.47729**	0.1706	0.12913	0.08384	0.0782	0.02215	0.39491	0.022	0.02673	0.02402	0.0221
	RP_N	**0.1355**	**0.7342**	**0.2248**	**0.22753**	**0.17306**	**0.1838**	**0.13163**	**0.5**	**0.156**	**0.15313**	**0.12442**	**0.10991**
	RT_N	0.02863	0.4738	0.1702	0.12787	0.08302	0.07762	**0.02222**	0.39499	0.0222	**0.0268**	0.02450	**0.02227**
	RPT_N	0.02431	0.35884	0.0792	0.08106	0.06026	0.04193	0.02177	0.33369	0.0262	0.0298	0.02628	0.02010
U*		UWOR testbed						UWR testbed					
Baseline	R	**0.03944**	**0.55786**	**0.28740**	0.20733	**0.13024**	**0.15292**	**0.03372**	0.46945	0.02780	0.03640	**0.03646**	**0.03501**
	RP	0.04711	0.67403	0.2358	0.224	0.17074	0.17017	0.03655	0.51417	0.1076	0.0972	0.07862	0.06876
	RT	**0.03945**	0.5585	**0.2884**	0.2078	**0.13038**	**0.15356**	0.03373	0.46972	0.0278	0.0364	0.03646	0.03504
	RPT	**0.04712**	**0.6747**	**0.2364**	**0.22427**	**0.17084**	**0.17043**	**0.03656**	**0.51445**	**0.1076**	**0.0972**	**0.07862**	**0.06877**
1%	R_N	0.03943	0.55738	0.2872	0.20733	0.13014	0.1528	0.03367	**0.47033**	**0.028**	**0.03646**	0.03630	0.03497
	RP_N	**0.13357**	**0.70045**	**0.2358**	**0.22627**	**0.1731**	**0.17797**	**0.1279**	**0.54195**	**0.1472**	**0.15453**	**0.13478**	**0.11312**
	RT_N	0.03927	**0.56499**	0.2838	**0.2082**	0.13022	0.15027	**0.03427**	**0.48844**	**0.03**	**0.0396**	**0.0401**	**0.03792**
	RPT_N	0.03656	0.4928	0.1128	0.12946	0.10548	0.08264	0.03431	0.45798	0.029	0.0454	0.04472	0.0394

testbed). This occurred as the noisy reputable and non-relevant documents confirm the concentration of selection decision on these reputable peers. Prioritising those non-relevant and reputable documents at the top of the results do not have any effect on the effectiveness of search result as these documents possess a little amount of reputation relevance scores based on the past interaction of users who highly choose the relevant documents. On the other hand, R_N and RT_N approaches have better retrieval results in few cases especially the testbeds with replication. In comparison with two approaches themselves, the results of R_N approach are much better than the RT_N approach. In addition, R_N and RP_N approaches reveal high retrieval results than RT_N and RPT_N comparing to their corresponding methods in baseline approaches. The reason is that the RT_N and RPT_N approaches take the content-based information of documents into consideration when they estimate the reputation score of the peers; which means that the reputable and non-relevant documents might have much enough content-based information to bias the quality of search result as shown

for RPT_N approach on all the cases.

In further noisy data as shown in Table B.19, the noisy approaches still have better and competitive results, especially in DL and U environments in some cases due to a large number of documents on average on each peer, while competitive results for ASIS environment that have small peers on a number of documents. In summary, we can see that even on a small proportion of noisy information provided by the users, the reputation-based approaches are robust and can give us competitive and better performance. These results confirm the research question (**RQ-6.3**).

6.4.4 Reputation-based and CORI Approach

The selection decision of the RT and RPT approaches built upon the query terms' term frequencies gathered from the reputable documents at the super-peer level. In order to improve these two simple approaches, I use the CORI resource selection method as content statistical information and combine its peers' scores to the R and RP reputation-based approaches as discussed in Subsection 6.2.3.

Table 6.7: Retrieval effectiveness α R + (1-α) CORI at 10% of Selected Peers

DL*	DLWOR testbed						DLWR testbed					
Method	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
RT	0.02612	0.44371	0.122	0.11866	0.0804	0.06805	0.02415	0.41621	0.028	0.028	0.03120	0.026
RPT	0.03786	0.59723	0.276	0.26933	0.1829	0.1998	0.03191	**0.52681**	0.186	0.18633	0.1362	0.13176
RCORI	**0.028**	**0.47669**	**0.181**$^{\blacktriangle}$	**0.143**$^{\blacktriangle}$	**0.09**$^{\triangle}$	**0.09221**$^{\blacktriangle}$	**0.02637**$^{\triangle}$	**0.46401**	**0.031**	**0.03066**	**0.0356**	**0.03139**$^{\blacktriangle}$
RPCORI	**0.17877**$^{\blacktriangle}$	**0.60588**	**0.398**$^{\blacktriangle}$	**0.31567**$^{\blacktriangle}$	**0.204**$^{\triangle}$	**0.2512**$^{\blacktriangle}$	**0.17889**$^{\blacktriangle}$	0.46518$^{\triangledown}$	**0.273**$^{\blacktriangle}$	**0.26167**$^{\blacktriangle}$	**0.1696**$^{\blacktriangle}$	**0.16689**$^{\blacktriangle}$
ASIS*	ASISWOR testbed						ASISWR testbed					
RT	0.02842	0.47268	**0.175**	**0.13**	0.08290	0.07797	0.02226	0.39705	0.022	0.027	0.0243	0.02238
RPT	0.04475	**0.72529**	0.272	0.265	**0.1994**	**0.2202**	0.03271	**0.56046**	0.138	0.13166	0.1078	0.10027
RCORI	**0.02932**	**0.47507**	0.172	0.12966	**0.0841**	**0.0802**	**0.02398**$^{\triangle}$	**0.41486**	**0.023**	**0.02800**	**0.0267**$^{\blacktriangle}$	**0.02456**$^{\blacktriangle}$
RPCORI	**0.15583**$^{\blacktriangle}$	0.69693	**0.304**$^{\triangle}$	**0.27367**	0.1882$^{\triangledown}$	0.21046$^{\blacktriangledown}$	**0.1489**$^{\blacktriangle}$	0.44548$^{\blacktriangledown}$	**0.188**$^{\blacktriangle}$	**0.185**$^{\blacktriangle}$	**0.1408**$^{\blacktriangle}$	**0.12169**$^{\blacktriangle}$
U*	UWOR testbed						UWR testbed					
RT	0.03297	0.37539	0.271	0.18233	0.1041	0.11258	0.02895	0.346	0.029	0.035	0.0316	0.02385
RPT	0.03733	0.44160	0.231	0.20367	0.1364	0.11274	0.03107	0.38095	0.114	0.09767	0.0728	0.05297
RCORI	**0.03822**$^{\blacktriangle}$	**0.53666**$^{\blacktriangle}$	**0.3**	**0.21233**$^{\blacktriangle}$	**0.1337**$^{\blacktriangle}$	**0.1527**$^{\blacktriangle}$	**0.03422**$^{\blacktriangle}$	**0.44446**$^{\blacktriangle}$	**0.047**$^{\blacktriangle}$	**0.065**$^{\blacktriangle}$	**0.0663**$^{\blacktriangle}$	**0.05123**$^{\blacktriangle}$
RPCORI	**0.19668**$^{\blacktriangle}$	**0.5907**$^{\blacktriangle}$	**0.414**$^{\blacktriangle}$	**0.34334**$^{\blacktriangle}$	**0.2121**$^{\blacktriangle}$	**0.2501**$^{\blacktriangle}$	**0.19104**$^{\blacktriangle}$	**0.38732**	**0.239**$^{\blacktriangle}$	**0.25333**$^{\blacktriangle}$	**0.1644**$^{\blacktriangle}$	**0.14876**$^{\blacktriangle}$

Table 6.7 shows the results of RCORI and RPCORI approaches in comparison to their corresponding RT and RPT baseline approaches at 10% of selected peers. The RCORI and RPCORI significantly outperform the baseline methods in almost over all the cases, especially in DL and U environments. However, although combining CORI approach reveals significant and competitive retrieval

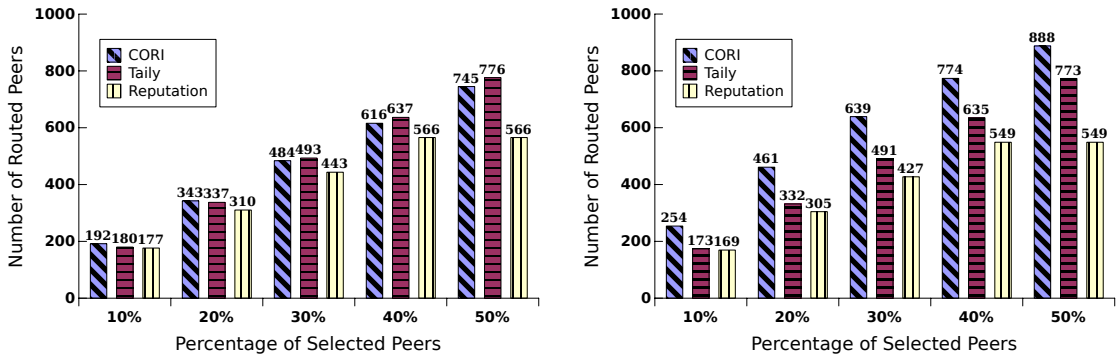
results, the simple RT and RPT approaches obtain more benefits. The reason is that the CORI approach requires more statistical information than the RT and RPT from single peer and collection of peers.

In summary, the CORI approach improves the retrieval effectiveness over all percentage of selections as also shown in Tables B.16, B.17, and B.18. Since we can infer that CORI approach has an ability to improve the results more and could be a solution to route a query to most relevant peers, this leads to answer the research question (**RQ-6.4**).

6.4.5 Network Traffic Efficiency

I measure the network efficiency by the number of messages needed to route a given query to the relevant peers; which is called message complexity and explains how efficient the proposed techniques are on semi-structured P2P-IR systems (Klampanos and Jose, 2007). However, in the architecture, the number of messages between super-peers are the same for the proposed and the baseline approaches. So I measure the network traffic as the number of messages between the super-peer and its peers. In the evaluation, I conducted the network efficiency at specific percentages between 10% and 50% of the ranked peers list at each super-peer. At each percentage of the selection, the super-peer determines a small number of candidate peers to route the given query to. Hence, each super-peer has a different number of selected peers due to a variety of peers at each super-peer. Even though I route a certain percentage of peers (10% to 50%) for all methods, I select those peers with non-zero scores only. This means from method to method the number of peers chosen vary for the same percentage of the selection. So, I differentiate between the percentage of *selected* peers represented in x-axis and number of *routed* peers represented in y-axis. In addition, all reputation-based methods have the same message complexity at a selected percentage level and hence I used one of them and called the Reputation method. Finally, the results were averaged over all testing queries (i.e, 100 topics) at each percentage of the reputation-based and other two baseline methods.

In DLWOR testbeds, Figure 6.2(a) shows the result of message complexity for the reputation-based methods and the other two baseline methods (CORI and Taily). The reputation-based approaches as shown in the figures below have con-



(a) DLWOR Test-bed.

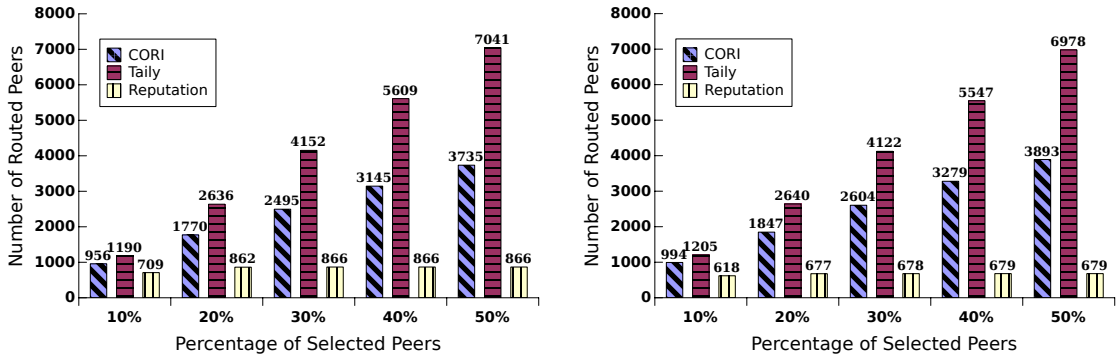
(b) DLWR Test-bed.

Figure 6.2: The efficiency of Reputation-based method on Digital Libraries

sistently and significantly[1] fewer numbers of messages on average 11.7% (8% to 24%) improved over the CORI approach and 11.6% (2% to 27%) on the Taily approach. In addition, as the percentage of selection peers increases, we can see the number of messages stabilises for the reputation-based approaches between 40% and 50% of selected peers. In user data simulation, a small number of peers with a large number of relevant documents are tagged implicitly as reputable peers. This skews the given query to those peers, which shows a power-law pattern of accessing peers as in real-life P2P scenarios. In the replication environment of DL^* family, the results in Figure 6.2(b) show that the proposed approaches outperform the other two baseline methods in reducing the number of messages on average 33.4% (33% to 38%) of routed peers over the CORI method and 13.2% (2% to 29%) of routed peers on the Taily approach with stable values between 40% and 50% of selected peers as in DLWOR testbed. The proposed approaches have a significantly fewer numbers of routed peers in DLWR testbed more than DLWOR testbed approximately on average of 3.4% (2% to 5%) of routed peers. As the replicated relevant documents over peers increase the probability of selecting reputable peers with highly relevant documents.

In $ASIS^*$ environments, the reputation-based approaches exhibits a significant reduction in a number of messages comparing with the CORI and the Taily approaches. Figure 6.3(a) presents the results of the approaches on ASISWOR

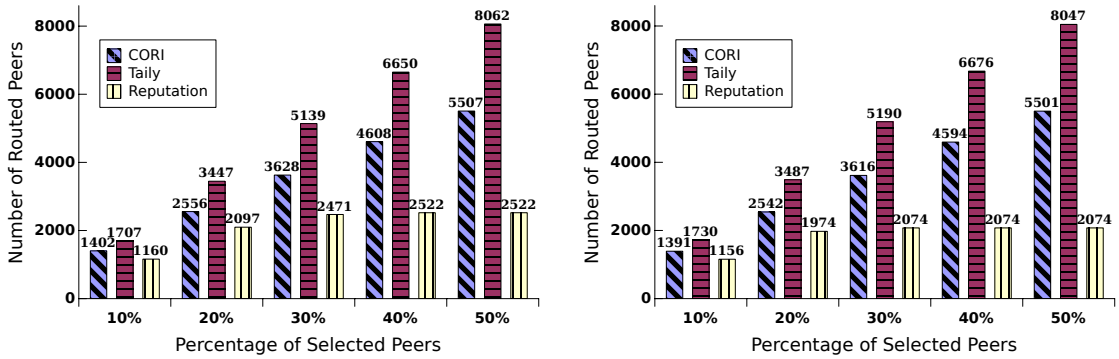
[1]The efficiency of methods based on a number of peers respond to the given query. The improved results are statistically significant at $p \leq 0.01$ using t-test method.



(a) ASISWOR Test-bed. (b) ASISWR Test-bed.

Figure 6.3: The efficiency of Reputation-based method on File Sharing

testbed that have significantly a small number of routed peers on average 58.42% (26% to 77%) for CORI approach and 71.86% (40% to 88%) of routed peers on Taily method. In ASISWR testbed in Figure 6.3(b), the approaches get better efficiency in reduced number of routed peers as on average 67.4% (38% to 83%) in comparison with the CORI approach and 77% (49% to 90%) on the Taily method. The two figures show stable values approximately between 20% and 30% of the selected peers in two testbeds. In addition, the result is much better in $ASIS^*$ family due to a large number of peers in this environment where the relevant documents are distributed on a small number of peers (Klampanos et al., 2005). Such distribution gives high-performance gain to the proposed methods compared to the two baseline methods.



(a) UWOR Test-bed. (b) UWR Test-bed.

Figure 6.4: The efficiency of Reputation-based method Uniform Distributed Systems

The efficiency of the proposed method on the uniform environment without

replication is shown in Figure 6.4(a). The proposed approaches obtain a significant reduction in a number of routed peers in comparison with the CORI and the Taily approaches. The results show the comparison with the two baseline methods which are on average 33.32% (17.3% to 54%) for the CORI method and 50.86% (32% to 69%) of routed peers on the Taily approach. The efficiency results of UWR testbed appear in Figure 6.4(b), which shows a reduction in a number of routed peers using the approaches compared with the CORI approach of 39.76% (17% to 62.3%) and the Taily method of 55.92% (33.2% to 74%). In both U^* testbeds, approximately between 30% and 50% of selected peers, we can see the values stabilise for the reputable method. Furthermore, we can see, the message complexity in $ASIS^*$ environment is distinctly improved more than U^* family; which are comparable as they have the same number of peers, with approximately 59% (39% to 66%) of routed peers between ASISWOR and UWOR testbeds and 62.8% (47% to 67%) between ASISWR and UWR testbeds.

In summary, the proposed reputation approaches in testbeds with replication have more reduction in a number of messages than the other without replication. Even in without replication scenarios, the methods perform consistently better. It is also important to note that the proposed approaches stabilise around 30% of selected peers. These results demonstrate, in addition to significant improvements in effectiveness, the proposed methods perform significantly or consistently better in reducing the message costs. This lead us to answer the research question **RQ-6.5** that addresses the ability of the proposed methods in reducing the routing messages in the semi-structured P2P-IR networks.

6.5 Reputation-based Query Routing Limitations

The reputation-based query routing approaches have the ability to route a query efficiently and significantly to the relevant peers in more effective quality results. In spite of performance and simplicity using the proposed approaches, a set of weaknesses unveils the challenges in applying them in real-life scenarios of P2P-IR systems. Three limitations might have influential effects in exploiting the reputation-based approaches for query routing scheme that can be faced to alleviate the problem, which are explained as follows:

1. In general, P2P-IR networks lack to actual testbeds that cover the realistic scenarios include: (i) dynamic updating of peers' collection during document downloading (i.e, real-time updating of documents in the systems); (ii) large size of peers' collections especially in digital library environments; (iii) using relevance feedback information on retrieved documents, either if its implicit, explicit, or pseudo- feedback, to reformulate the given query; (iv) and the dynamic nature of users in the network, which is called churn rate (Stutzbach and Rejaie, 2006), where a user can join and leave the network at any time. A set of testbeds have been emerged (Lu and Callan, 2003; Klampanos et al., 2005), but still, they are not sufficient to mimic such behaviours on various scenarios. However, since the reputation-based approaches inherently depend on the users' implicit feedback provided by clicking through retrieved documents in the final result list, the implicit feedback information in P2P-IR networks reveals transmission costs to sent back such information to the responsible super-peer. Therefore, although I have an evaluation experiments, including the implicit feedback behaviour through using the judgement assessment file, the implicit feedback approaches need to be further studied in more detail for enhancing the reputation-based approaches including the user-oriented methodology as a real user behaviour.

2. The approaches use the reputation-based data structure as an inverted index of a super-peers' peers and their reputable relevant and/or non-relevant documents. This data structure at super-peer level needs more storage in the case of increasing number of reputable documents during the interaction sessions in the system. This large amount of storages manifests a challenge in processing a query, especially with various terms at the majority of requests and updating new reputable documents. In terms of processing a query, P2P-IR networks end up with more computation costs, especially in the RT and RPT methods that require estimating the query terms statistical information from the reputable documents. To overcome such challenge, a caching-based technique is required to alleviate the processing computation demands in indexing the reputable documents.

3. The approaches, especially the R and RP methods estimate the reputation

score of a peer by aggregating their documents reputation values using the reputation-based data structure at the peer's super-peer. This reputation score is not sufficiently accurate due to the summation mechanism that hides much adequate information behind. For instance, a peer shares a small number of documents with various topics to the system. During building the reputation value of a document associated with that peer, the implicit feedback information is, in the past interaction sessions, assigned for various training queries of various topics. Since, given a run-time query on a specific topic, the peer's reputation score is less accurately aggregated on different reputable document scores with various topics and deviates the score from accurate value.

6.6 Conclusions

In this chapter, I proposed reputation-based query routing approaches in semi-structured P2P-IR network that exploit past users' interaction data to route a query to reputable peers of relevant documents. I built four routing methods; which are R, RP, RT, and RPT as well as combining CORI approach as a content-based statistical information to R and RP instead of RT and RPT approaches. In order to validate the robustness of the proposed approaches, I conduct experiments on different scenarios. Scenarios 1 and 2 (i.e, leave-one-out method) reflect of using the past data information in the estimation of reputable peers, scenario 3 studies other various leave-out methods, and scenario 4 analyses the robust of the proposed approaches under noisy usage information (reputable of non-relevant documents).

However, in particular, as shown in Section 6.4, the reputation-based measures show statistically significant improvement in comparison to the state-of-the-art CORI and Taily based selection algorithms in P2P-IR scenarios. In addition, this reduces the message complexity significantly. I have conducted the experiments on six different testbeds mimicking various real-life P2P scenarios. Since real-life P2P networks show a power-law pattern in accessing documents and peers, the proposed approach based on reputation exploits this behaviour, I can reasonably conclude the feasibility of the approach for real-life scenarios. Given that, sig-

nificant performance improvements, especially in scenario 1 and 2, in retrieval effectiveness are seen in P2P systems using reputation values and hence the research question **RQ-6.1** is answered affirmatively to validate the idea of using past interaction data to enhance the query routing and subsequently the retrieval effectiveness. In addition, I did some experiments to evaluate the robustness of the proposed methods through changing the training and testing information in scenario 3. The results show a better performance of the proposed approaches, even with a lack of usage information to calculate the reputation scores for the peers, which lead us to answer the research question **RQ-6.2** that addresses using variants amount of interaction data like 25%, 50%, and 70% to calculate the reputation values of peers at query run-time. The users in the system might click and download the non-relevant documents as discussed in scenario 4, which in these cases could affect the retrieval effectiveness. But as shown from the results even given this information of noisy, the proposed approaches still obtain significant and competitive results that answer the research question **RQ-6.3** that evaluates using non-relevant documents in interaction data simulation for reputation-based query routing and evaluates its effect on retrieval quality. I also discussed the effectiveness of combining the CORI approach as a complicated content-based statistical information instead of term frequencies of query terms. The results exhibited more significant improvements in almost all the cases of measurement metrics in three environments (DL, ASIS, and U), which leads us to answer the research question **RQ-6.4** that evaluates the applicability of using the CORI score as statistical content information combined with reputation value of peers for enhancing the query routing and retrieval effectiveness. Subsequently, the number of selected peers has an effect on the time taken to retrieve the final results in the system. In other words, a higher number of peers used increases the time taken to execute queries on the distributed indices (i.e. peer's collection) and to merge the results list at a super-peer level. Through this analysis, I have clearly shown that the proposed approaches of reputation-based P2P retrieval reduce the network traffic significantly compared to the baseline CORI and Taily selection approaches, which answers the research question **RQ-6.5** that addresses the evaluation of reputation-based query routing methods in message complexity which is measured as a number of routed peers.

Part IV

Take-away Messages and Future Research

In the final part of this thesis, I summarise the various works presented in the previous chapters on query routing in co-operative semi-structured P2P-IR networks. In particular, I focus on the high level "take away" messages which can be drawn from this work as well as dwell on new areas for future research.

Chapter 7

Conclusions and Future Work

> "A conclusion is the place where you got tired thinking."
>
> — Martin H. Fischer, (1879 - 1962)

7.1 Introduction

In this chapter, I present summaries of previous chapters and explain the important results and messages. The conclusions gear the spotlight toward the importance of study and highlight the key points in enhancing and improving the retrieval performance in semi-structured P2P-IR networks. During this study, I faced a set of future directions that are sparked in each contribution chapter and have to be discussed as motivations for new researchers. I will expose these directions as further research works in the future.

7.2 Conclusions

This thesis started with an introduction as a blueprint for the remaining chapters, including the scope of the research study, problem statement, challenges, and contributions of work with high-level research questions. I also navigated over a survey of literature review including information retrieval systems and distributed information retrieval paradigm. The conclusions section illustrates and identifies only the summary of contribution chapters to support and promote the objectives behind this thesis. In order to simplify the conclusions, I will separately introduce

each contribution chapter along with its results and an open key for the next chapter.

7.2.1 Semi-structured Peer to Peer Information Retrieval

In this chapter, I built cluster-based semi-structured P2P-IR architectures using cluster hypothesis to group the peers with similar domain under a specific super-peer. These cluster-based architectures have two types of peers; super-peers and normal peers. The process essentially passed through two stages which are clustering in-peer documents and global network clustering. In-peer clustering uses Bisecting K-means algorithm to group the documents in each peer to determine its topics. This stage of clustering explores and identifies the peer's information tendency and interests. In the subsequent stage, the cluster centroids created in the peers were used as input for the further clustering process to build the super-peer networks. In order to build different super-peer network topologies, I constructed three cluster-based topological models of the semi-structured P2P-IR networks; K-means, Half K-means-Single-Pass, and Approximation single-Pass models. The K-means model uses the peers' centroids as input to k-means clustering algorithm to create 50 initial attractive cluster points each of which assigned to one super-peer. Half K-means Single-Pass model uses the half cluster centroids of the peers to construct the super-peers using K-means model, then the single pass clustering algorithm is used to assign the remaining centroids to the created super-peers through estimating the cosine similarity between these centroids and the super-peer centroids. Finally, in Approximation single-Pass model, I distributed the peers' centroids into 8 packets and at each packet I created a set of centroids that are aggregated over all packets to construct 50 description centroids to be handled by different super-peers. The research question based on this part was (i.e, **RQ-4.1**) on the performance of different clustering approaches in semi-structured P2P-IR networks. The retrieval results using these topologies on evaluating testbeds addressed such question and showed different retrieval performance between architectures where k-means model on evaluating testbeds outperformed the other two models. The K-means model was used as a target model for studying a set of parameters and for the next two chapters.

However, a set of experimental parameter settings was examined and studied in semi-structured P2P-IR network by using K-means model as a target network to answer a set of important research questions. These parameters include the effectiveness of increasing number of super-peers, the impact of using different retrieval models on each peer, and the scalability and robustness of semi-structured P2P-IR networks. The experiments showed a set of results that is mentioned here along with addressed research questions, which are: (i) using conventional information retrieval models have an effect where there are retrieval models perform better in distributed systems such as LGD model than the other models in comparison with the centralised systems (**RQ-4.2**), (ii) the small number of super-peers are closer to centralised systems with more effective retrieval quality (**RQ-4.3**), (iii) and the robustness of the semi-structured P2P-IR networks that have less effectiveness on departure or failure peers (**RQ-4.4**).

In summary, semi-structured P2P-IR networks are effective and sufficient as a promising technology for information retrieval tasks. The question is how to improve the query routing under such networks and exploiting the coherence semantic groups to build an effective and efficient routing method. There have been several resource selection methods in the federated search, especially meta-search environment, and it would be more benefits if we deploy these techniques in semi-structured P2P-IR networks.

7.2.2 Cooperative Resource Selection Methods in Federated Search

This chapter discussed a set of query routing techniques that conducted under the k-means model. First, I exploited the cluster centroids at the super-peer level to build an inverted index of term and peers along with term scores; which is called Inverted PeerCluster Index (IPI) approach. This resource selection method adopts the conventional inverted index in the IR to the level of peer-clusters. The research question **RQ-5.1** examines the benefit of using the cluster centroids at super-peer level for resource selection method in semi-structured P2P-IR networks. An extensive analysis showed that the simple and effective IPI approach is seen to emulate the state-of-the-art resource selection methods designed for

meta-search environments in terms of retrieval effectiveness and message complexity. Second, I empirically benchmarked well-known document retrieval methods against the state-of-the-art resource selection methods designed for general P2P-IR. Based on the research question **RQ-5.2** that addresses the applicability of using the document retrieval models as resource selection methods, an extensive analysis of retrieval effectiveness over the target semi-structured P2P-IR network using classical IR evaluation metrics validated the hypothesis convincingly, with document retrieval methods consistently outperforming the others. This establishes that document retrieval methods could be the preferred choice for resource selection in clustered P2P-IR environments. Third, I proposed a Learning to Route (LTRo) approach to route a given query to the most likely relevant peers. I used the state-of-the-art Learning to Rank (LtR) algorithms to train a model and predict the peers' scores for future queries. The training set is built based on specific features of resource selection approaches along with a label that is assigned as a discrete value of the number of relevant documents at top 10 retrieved documents. By experimenting under different conditions and testbeds, I studied the performance of different LtR algorithms in predicting the testing query set as well as the retrieval effectiveness conducted on evaluation metrics of testing queries. The results answered the research question **RQ-5.3** that refers to the effectiveness and efficiency of using LtR algorithms as resource selection methods (i.e, LTRo) and showed the effectiveness of LTRo approach on testing set and an improvement in the routing processes in comparison with the state-of-the-art methods such as the Taily and classification-based resource selection algorithms.

In summary, I examined the retrieval effectiveness and efficiency of the proposed IPI model, adapted document retrieval methods and LTRo approach on the target k-means model of semi-structured P2P-IR network and also I analysed the document retrieval methods and LTRo approach on the federated meta-search environment. The results showed that these methods are preferable, and competitive as resource selection approaches for both environments. However, I did not investigate, in this chapter, the impact of user interference in providing feedback to the system on the retrieved documents to improve the retrieval performance for future queries. In a further study, I left such study to the next chapter where the k-means model is still the target network.

7.2.3 Reputation-based Query Routing

In the previous chapter, I studied the effectiveness and efficiency of the proposed IPI, adapted document retrieval, LTRo models, and other state-of-the-art methods on the K-means model of semi-structured P2P-IR architectures where the users are only consumer or information provider. The users' behaviours have an impact on security perspective in P2P networks; the users rate an object while the system aggregates this information for future decision-making. In this chapter, I simulated the user interaction data to provide implicit feedback from clicking through data (i.e, documents) to construct a reputation-based data structure at each super-peer for query routing. Under using these data structures, I proposed reputation-based query routing approaches to route a query to reputable peers of relevant documents. I built four routing methods; which are R, RP, RT, and RPT as well as combining CORI approach as a content-based statistical information to R and RP instead of RT and RPT approaches. I conducted experiments on different scenarios to validate the robustness of these approaches. Scenarios 1 and 2 (i.e, leave-one-out method) reflect using the past data information in estimating reputable peers. I answered the research question **RQ-5.1** that evaluates the benefit of exploiting the whole or part (i.e, leave-one-out method) of interaction data at a query run-time to assign reputation scores to the peers where the results showed competitive and significant retrieval quality compared to the Taily and CORI baseline methods in both scenarios. Scenario 3 studies other various leave-out methods to build reputation scores for the peers, which are (25-75)%, (50-50)%, and (75-25)% methods to answer the research question **RQ-5.2** that refers to the effectiveness of the reputation-based query routing methods on varying amount of interaction data. The results showed even under such varying data a better performance compared to the baseline methods. Scenario 4 analyses the robustness of the approaches under noisy usage information (reputable of non-relevant documents). The results proved the research question **RQ-5.3** that examines the effect of noisy information by providing implicit feedback on non-relevant documents where reputation-based query routing methods under noisy information can give competitive and better results. Finally, in terms of effectiveness, I examined using the CORI score as statistical information on

reputation-based methods (i.e, RCORI and RPCORI) that answered the research question **RQ-5.4** of the effect using CORI method as content-based statistical information where RCORI and RPCORI methods improve the retrieval results. Since real-life P2P networks show a power-law pattern in accessing documents and peers, the proposed reputation-based approaches exploit this behaviour which can reasonably conclude the feasibility of the approaches for real-life scenarios. Using extensive experimental evaluation, I proved that the reputation-based measures have statistically significant improvement in comparison to the state-of-the-art CORI and Taily based selection algorithms in the P2P-IR scenarios. In addition, this reduces the message complexity significantly as an answer to the research question **RQ-5.5** that states the efficiency of using reputation-based query routing methods in semi-structured P2P-IR networks.

In summary, I showed how to exploit the feedback from the users to enhance the query routing quality. Hence users' behaviours have a crucial effect on routing the query to improve the retrieval effectiveness, which are approximately more than 56% in almost all the cases. In addition, the message complexity was significantly decreased on average less than 50% of a number of routed messages.

7.3 Future Work

There are several works not covered in this thesis due to space and time limitations. Therefore, I explain a set of future works for further research works and to be as directions for new researchers. As I discussed the conclusion of each chapter separately, I will follow the idea for future works as follows.

7.3.1 Semi-structured Peer to Peer Information Retrieval

The cluster hypothesis is used to build the semi-structured P2P-IR networks. The hypothesis depends on two levels of clustering. In this work, I used Bisecting K-means algorithm for inter-peer clustering that is chosen by efficiency consideration (Steinbach et al., 2000) as well as I used three different clustering models for building the network. In future works, I will examine other clustering algorithms selected from different families and study their computational and storage costs. The effective and efficient clustering algorithm is the one that can be used for a

large number of documents (scalable) with limited time (in second) and has to be used in real-time situations. This is due to the dynamic growth of documents in the network. Furthermore, I will exploit the classification-based approaches that are built under a large number of documents or corpus with different topics to construct a classifier that identifies the topics inside each peer. The classifier categorises the centroids in each peer into a set of topics (e.g, sport, health, art) and then the system connects each peer based on its topics to different super-peers that are previously determined and responsible for specific topics. In addition, I will expand this study in the used parameter settings such as using different retrieval models for peers in the network, examine a various number of super-peers and exhaust the system with more churn rate values to examining its robustness under this pressure. Finally, I will use the structured P2P networks to organise the centroids' terms of the super-peer level and hash them as a key value into DHT (Distributed Hash Table) data structure. This means that super-peer level is organised as a structured network while peers level as an unstructured network. I highlight research questions for future works (FW) related to this chapter as follows:

FW-4-RQ1: *What is the effect of using different clustering algorithms on building semi-structured P2P-IR networks and what are the computational and storage costs of using these algorithms?*

FW-4-RQ2: *What is the effectiveness and efficiency of using classification-based semi-structured P2P-IR networks in comparison with cluster-based networks?*

FW-4-RQ3: *Does exploiting the Distributed Hash Table (DHT) data structure that is used in structured P2P-IR networks to hash the centroids' terms at super-peer level enhance query routing quality?*

7.3.2 Cooperative Resource Selection Methods in Federated Search

I conducted experiments on the semi-structured P2P-IR networks and studied a set of resource selection methods from meta-search environment. I also examined

on the same network proposed methods such as IPI, adapted document retrieval and LTRo approaches. I plan to explore methods to adapt the IPI approach to uncooperative environments by devising sampling strategies that could identify intra-peer clusters so that the notion of peer clusters may extend naturally. Another promising direction is to leverage query logs as and when they accumulate, in improving IPI-based resource selection. From a document retrieval perspective, the results indicate a potential convergence of the document retrieval and resource selection tasks in the clustered P2P-IR architecture so that future advancements in document retrieval may be effectively leveraged to achieve corresponding gains in resource selection on the target framework. With the improved accuracy of document retrieval techniques being apparent from the study, their indexing overhead and messaging costs need to be subject to further analysis and study so that their uptake may be facilitated. Finally to the LTRo approach, I will use an untested feature set such as query topic, uncooperative single resource selection methods, and other peer-based features such as peer reputation (and trustworthiness). In addition, I will test other learning to rank algorithms on the environments of the different testbeds. The Future work research questions are as follows:

FW-5-RQ1: *What is the effect of using sampling documents techniques as in uncooperative environments and their corresponding resource selection techniques in improving the query routing effectiveness and efficiency on semi-structured P2P-IR networks?*

FW-5-RQ2: *Is it possible to exploit the inter-peer clusters rather than sampling documents on uncooperative for resource selection methods?*

FW-5-RQ3: *How to use query logs to improve the IPI approach where the high frequently used term in past interaction is updated with high value in IPI inverted index data structure?*

FW-5-RQ4: *What is the effect of using advanced retrieval models in the semi-structured P2P-IR network for retrieval quality? What are the best parameters for each document retrieval resource selection model?*

FW-5-RQ5: *What is the effect of studying each feature separately and other new features on improving the LTRo approach? What is the effect of using other LtR algorithms as resource selection methods?*

7.3.3 Reputation-based Query Routing

I simulate the user interaction in providing feedback on the retrieved documents on semi-structured P2P-IR networks. First, I generated the training query set and conducted an experiment to simulate the user interaction data through providing the users with a ranked document list where each user clicks or downloads specific documents that might relevant to his information need. The implicit feedback generated by the users is aggregated at super-peer level; each super-peer manage the information related to the documents in its peers. The aggregated feedback on a specific document can be seen as a reputable value that is used later to estimate the peer's reputation score for query routing. The future work can be summarised as follows: (i) expanding the boundaries of training and testing leave-out methods, (ii) incorporating the malicious behaviour of peers in P2P networks such as providing incorrect feedbacks to the system and studying their impacts on query routing quality, (iii) analysing the incentive and punishment techniques in P2P security aspect for improving query routing, (iv) studying more effective techniques to estimate the reputation values of documents based on given feedback, (v) and deploying the approach used in (Zhang, 2011) of incorporating the trustworthiness and relevance on the same documents to improve the security in the semi-structured P2P-IR network. The corresponding future research questions are as follow:

FW-6-RQ1: *In the experiments, I used four leave-out methods to examine the robustness of the reputation-based query routing approaches, but what is the effect of using other leave-out methods with varying boundaries to stress the reputation values of the peers in the system?*

FW-6-RQ2: *If we have malicious peers with many false provided feedback, how would they have an impact on the reputation-based approaches and the quality of query routing?*

FW-6-RQ3: *How can we use the incentive and punishment techniques to encourage selfish peers to provide feedback and prevent malicious peers from providing incorrect feedbacks and how can we benefit from these techniques to improve the query routing?*

FW-6-RQ4: *Using simple aggregated feedback from different users might have an effect because it depends on different factors such as the number of interactions of each user, the credibility of the user to provide that information and the number of interaction that occurred on that document. Therefore, how can we incorporate these factors and what are their effects on query routing? What is the effect of using the actual user feedback on reputation-based query routing methods?*

FW-6-RQ5: *How to incorporate the approach used in structured P2P-IR in (Zhang, 2011) to estimate the reputation values of documents and study the effectiveness and efficiency of that approach on the semi-structured P2P-IR networks?*

References

Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 6 - Volume 6*, HICSS '00, pages 6007–, Washington, DC, USA, 2000. IEEE Computer Society. ISBN 0-7695-0493-0. 12, 138, 139

Harry T. Yani Achsan and Wahyu Catur Wibowo. A fast distributed focused-web crawling. *Procedia Engineering*, 69:492 – 499, 2014. ISSN 1877-7058. doi: http://dx.doi.org/10.1016/j.proeng.2014.03.017. 18

Farooq Ahmad and Grzegorz Kondrak. Learning a spelling error model from search query logs. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing*, pages 955–962, Vancouver, BC, Canada, 2005. ACL. 20

Rami S. Alkhawaldeh and Joemon M. Jose. Experimental study on semi-structured peer-to-peer information retrieval network. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 3–14. 2015. doi: 10.1007/978-3-319-24027-5_1. 7, 72, 111, 113

Rami S. Alkhawaldeh, Joemon M. Jose, and Deepak P. Evaluating document retrieval methods for resource selection in clustered p2p ir. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2073–2076, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4073-1. doi: 10.1145/2983323.2983912. 122

Robin Aly, Djoerd Hiemstra, and Thomas Demeester. Taily: Shard selection using the tail of score distributions. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 673–682, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484033. 52, 53, 60, 118, 121, 129, 144

Gianni Amati. *Probability models for information retrieval based on Divergence From Randomness*. PhD thesis, University of Glasgow, 2003. 34, 35

Gianni Amati and C. J. van Rijsbergen. Term frequency normalization via Pareto distributions. In *Proceedings of the 24th European Conference on Information Retrieval*, pages 183–192, Glasgow, UK, 2002. Springer. ISBN 3-540-43343-0. 34, 35

Ioannis Anagnostopoulos and Ioannis Avraam. A comparison over focused web crawling strategies. *2012 16th Panhellenic Conference on Informatics*, 00(undefined):245–249, 2011. doi: doi.ieeecomputersociety.org/10.1109/PCI.2011.53. 18

Stephanos Androutsellis-Theotokis and Diomidis Spinellis. A survey of peer-to-peer content distribution technologies. *ACM Comput. Surv.*, 36(4):335–371, December 2004. ISSN 0360-0300. doi: 10.1145/1041680.1041681. 63, 65, 84

Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, 1(1): 2–43, 2001. ISSN 1533-5399. doi: 10.1145/383034.383035. 19

Jaime Arguello, Jamie Callan, and Fernando Diaz. Classification-based resource selection. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1277–1286, New York, NY, USA, 2009a. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646115. 52, 53, 61, 128

Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. Sources of evidence for vertical selection. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 315–322, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1571997. 61, 62

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Pearson Education Ltd., Harlow, UK, 2 edition, 2011. ISBN 978-0-321-41691-9. 2, 3, 15, 19, 23, 41, 46

Peter Bailey, Arjen P. De Vries, Nick Craswell, and Ian Soboroff. Overview of the trec-2007 enterprise track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC*, 2007. 4, 43, 51

F. Baskaya, H. Keskustalo, and K. Järvelin. Modeling behavioral factors in interactive information retrieval. In *Proc. 22nd ACM CIKM*, pages 2297–2302, 2013. 48

N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, (5):133–143, 1980. 43

Matthias Bender, Matthias Bender, Sebastian Michel, Sebastian Michel, Gerhard Weikum, Gerhard Weikum, Christian Zimmer, and Christian Zimmer. Bookmark-driven query routing in peer-to-peer web search. In *Proceedings of the SIGIR Workshop on Peer-to-Peer Information Retrieval. (2004) 46–57*, pages 46–57, 2004. 58

Michael Bendersky, Donald Metzler, and W. Bruce Croft. Learning concept importance using a weighted dependence model. In Brian D. Davison 0001, Torsten Suel, Nick Craswell, and Bing Liu 0001, editors, *WSDM*, pages 31–40. ACM, 2010. ISBN 978-1-60558-889-6. 40

Michael Bendersky, W. Bruce Croft, and Yanlei Diao. Quality-biased ranking of web documents. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 95–104, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935849. 20

Michael K. Bergman. The deep web: Surfacing hidden value. *Journal of Electronic Publishing*, 7(1), 2001. 2, 3, 18, 48

Pia Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003. 41, 43

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. 38, 39

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V. 20

Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002. ISSN 0163-5840. doi: 10.1145/792550.792552. 1

Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 25–32, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009000. 90

Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Inf. Retr.*, 10(6):491–508, December 2007. ISSN 1386-4564. doi: 10.1007/s10791-007-9032-x. 42

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, Bonn, Germany, 2005. ACM. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102363. 39, 45

Christopher J.C. Burges. From ranknet to lambdarank to lambdamart: An overview. Technical Report MSR-TR-2010-82, June 2010. 40

C.J.C. Burges, R. Ragno, and Q.V. Le. Learning to rank with non-smooth cost functions. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, January 2007. 38, 40

James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In *Proc. of SIGIR*, pages 21–28, New York, NY, USA, 1995. ACM. ISBN 0-89791-714-6. doi: 10.1145/215206.215328. 52, 55, 56, 61, 62, 118, 128, 144

Jamie Callan. Distributed information retrieval. In *In: Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000. 52, 53, 55, 110, 118

Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, April 2001. ISSN 1046-8188. doi: 10.1145/382979.383040. 52, 57

Jamie Callan, Margaret Connell, and Aiqun Du. Automatic discovery of language models for text databases. *SIGMOD Rec.*, 28(2):479–490, June 1999. ISSN 0163-5808. doi: 10.1145/304181.304224. 57

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. Technical Report MSR-TR-2007-40, Microsoft Research, April 2007. 37, 39

Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, January 2012. ISSN 0360-0300. doi: 10.1145/2071389.2071390. 20

Ben Carterette and Rosie Jones. Evaluating web search engines using clickthrough data. In *In Proceedings of NIPS 2007*, 2007. 150

Vitor R. Carvalho, Matthew Lease, and Emine Yilmaz. Crowdsourcing for search evaluation. *SIGIR Forum*, 44(2):17–22, January 2011. ISSN 0163-5840. doi: 10.1145/1924475.1924481. 42

Carlos Castillo. *Effective web crawling*. PhD thesis, University of Chile, 2004. 17

Suleyman Cetintas, Luo Si, and Hao Yuan. Learning from past queries for resource selection. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1867–1870, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646251. 52, 61

Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the Eighth International Conference on World Wide Web*, WWW '99, pages 1623–1640, New York, NY, USA, 1999. Elsevier North-Holland, Inc. 18

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646033. 46, 129

Hanhua Chen, Hai Jin, Xucheng Luo, Yunhao Liu, and L.M. Ni. Bloomcast: Efficient full-text retrieval over unstructured p2ps with guaranteed recall. In *Cluster Computing and the Grid, 2009. CCGRID '09. 9th IEEE/ACM International Symposium on*, pages 52–59, 2009. doi: 10.1109/CCGRID.2009.50. 71

Hanhua Chen, Hai Jin, Xucheng Luo, Yunhao Liu, Tao Gu, K. Chen, and L.M. Ni. Bloomcast: Efficient and effective full-text retrieval in unstructured p2p networks. *Parallel and Distributed Systems, IEEE Transactions on*, 23(2):232 –241, feb. 2012. ISSN 1045-9219. doi: 10.1109/TPDS.2011.168. 64, 77

Mu-Song Chen and Michael T. Manry. Conventional modeling of the multilayer perceptron using polynomial basis functions. *IEEE Trans. Neural Networks*, 4(1):164–166, 1993. doi: 10.1109/72.182712. 127

Ruichuan Chen, Xuan Zhao, Liyong Tang, Jianbin Hu, and Zhong Chen. Cuboidtrust: a global reputation-based trust model in peer-to-peer networks. In *Proceedings of the 4th international conference on Autonomic and Trusted Computing*, ATC'07, pages 203–215, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-73546-1, 978-3-540-73546-5. 78

Sergey Chernov, Pavel Serdyukov, Matthias Bender, Sebastian Michel, Gerhard Weikum, and Christian Zimmer. Database selection and result merging in p2p web search. In *Proc. of DBISP2P'05/06*, pages 26–37, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-71660-0. 58, 70

C. W. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. *ASLIB Cranfield Research Project*, 2, 1966. 41

Cyril W. Cleverdon. On the inverse relationship of recall and precision. *Journal of Documentation*, 28(3):195–201, 1972. doi: 10.1108/eb026538. 44

Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17 (4):32–38, July 2013. ISSN 1089-7801. doi: 10.1109/MIC.2012.95. 42

William S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971. doi: http://dx.doi.org/10.1016/0020-0271(71)90024-6. 25

Gordon V. Cormack, Mark D. Smucker, and Charles L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.*, 14(5):441–465, October 2011. ISSN 1386-4564. doi: 10.1007/s10791-011-9162-z. 21

David Cossock and Tong Zhang. *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings*, chapter Subset Ranking Using Regression, pages 605–619. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-35296-9. doi: 10.1007/11776420_44. 38

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 87–94, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. doi: 10.1145/1341531.1341545. 45

Fabio Crestani and Ilya Markov. Distributed information retrieval and applications. In *Proceedings of ECIR*, ECIR'2013, pages 865–868. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-36972-8. doi: 10.1007/978-3-642-36973-5_104. 52, 54, 110

Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009. ISBN 0136072240, 9780136072249. 2, 16, 18, 21, 41, 46

Francisco Matias Cuenca-Acuna, Christopher Peery, Richard P. Martin, and Thu D. Nguyen. Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities. In *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing*, HPDC '03, pages 236–, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-1965-2. 70, 73

Ronan Cummins. A study of retrieval models for long documents and queries in information retrieval. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 795–805, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883009. 103

Jeffrey Dean. Challenges in building large-scale information retrieval systems. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, page 1, Barcelona, Spain, 2009. ACM. ISBN 978-1-60558-390-7. doi: 10.1145/1498759.1498761. 3, 19

Thomas Demeester, D Trieschnigg, D Nguyen, and D Hiemstra. Overview of the trec 2013 federated web search track. In *Proceedings of the Text Retrieval Conference*, pages 1–11, 2013. 113

Bekir Taner Dinçer. IRRA at TREC 2012: Divergence from independence (DFI). In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, 2012. 36

Christos Doulkeridis, Akrivi Vlachou, Kjetil Nø rvåg, Yannis Kotidis, and Michalis Vazirgiannis. Efficient search based on content similarity over self-organizing p2p networks. *Peer-to-Peer Networking and Applications*, 3:67–79, 2010. ISSN 1936-6442. 10.1007/s12083-009-0058-2. 84

Yi Fang, Naveen Somasundaram, Luo Si, Jeongwoo Ko, and Aditya P. Mathur. Analysis of an expert search query log. In *Proceedings of SIGIR '11*, pages 1189–1190, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010113. 2

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968. ISBN 0471257087. 36

Nicola Ferro. CLEF 15th birthday: Past, present, and future. *SIGIR Forum*, 48(2): 31–55, 2014. doi: 10.1145/2701583.2701587. 42

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A*, 222(594–604):309–368, 1922. doi: 10.1098/rsta.1922.0009. 33

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. 41

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, December 2003. ISSN 1532-4435. 41

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000. 38, 39

Fredric C. Gey, Noriko Kando, and Carol Peters. Cross-language information retrieval: the way ahead. *Inf. Process. Manage.*, 41(3):415–431, 2005. 42

William Goffman. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2(2):73–78, 1964. 15

Dion Goh, Dion Goh, and Schubert Foo. *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2007. ISBN 1599045435, 9781599045436. 79

J. L. Goldberg. Cdm: An approach to learning in text categorization. In *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, TAI '95, pages 258–, Washington, DC, USA, 1995. IEEE Computer Society. ISBN 0-8186-7312-5. 56

Félix Gómez Mármol, Gregorio Martínez Pérez, and AntonioF. Gómez Skarmeta. Tacs, a trust model for p2p networks. *Wireless Personal Communications*, 51(1):153–164, 2009. ISSN 0929-6212. doi: 10.1007/s11277-008-9596-9. 78

Gregory Goth. Degrees of separation. *Commun. ACM*, 55(7):13–15, July 2012. ISSN 0001-0782. doi: 10.1145/2209249.2209255. URL http://doi.acm.org/10.1145/2209249.2209255. 70

Luis Gravano and Hector Garcia-Molina. Generalizing gloss to vector-space databases and broker hierarchies. In *Proceedings of the 21th International Conference on Very Large Data Bases*, VLDB '95, pages 78–89, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-379-4. 55

Luis Gravano, Héctor García-Molina, and Anthony Tomasic. The effectiveness of gioss for the text database discovery problem. *SIGMOD Rec.*, 23(2):126–137, May 1994. ISSN 0163-5808. doi: 10.1145/191843.191869. 54

Luis Gravano, Héctor García-Molina, and Anthony Tomasic. Gloss: Text-source discovery over the internet. *ACM Trans. Database Syst.*, 24(2):229–264, June 1999. ISSN 0362-5915. doi: 10.1145/320248.320252. 52, 55, 118

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–274, Boston, MA, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: http://doi.acm.org/10.1145/1571941.1571989. 20

Mo Hai and Shuhang Guo. A distributed node clustering mechanism in p2p networks. In Longbing Cao, Jiang Zhong, and Yong Feng, editors, *Advanced Data Mining and Applications*, volume 6441 of *Lecture Notes in Computer Science*, pages 553–560. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-17312-7. doi: 10.1007/978-3-642-17313-4_57. 84

Donna Harman. Overview of the second Text REtrieval Conference (TREC-2). In *Proceedings of the 2nd Text REtrieval Conference*, Gaithersburg, MD, USA, 1993. 44

S. P. Harter. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *Journal of the American Society for Informaiton Science*, 26(4):197–206, 1975. 27, 34

David Hawking. Challenges in enterprise search. In *Proceedings of the 15th Australasian Database Conference - Volume 27*, ADC '04, pages 15–24, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc. URL http://dl.acm.org/citation.cfm?id=1012294.1012297. 4, 51

Ben HE and Iadh Ounis. A study of parameter tuning for term frequency normalization. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 10–16, New York, NY, USA, 2003. ACM. ISBN 1-58113-723-0. doi: 10.1145/956863.956867. URL http://doi.acm.org/10.1145/956863.956867. 37

Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, and Nirav Shah. Crawling deep web entity pages. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 355–364, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3. doi: 10.1145/2433396.2433442. 2, 48

Djoerd Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001. 33

Dzung Hong, Luo Si, Paul Bracke, Michael Witt, and Tim Juchcinski. A joint probabilistic classification model for resource selection. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 98–105, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835468. 52, 61, 128

Chuan Huang, Yinzi Chen, Wendong Wang, Yidong Cui, Hao Wang, and Nan Du. A novel social search model based on trust and popularity. In *Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on*, pages 1030–1034, 2010. doi: 10.1109/ICBNMT.2010.5705245. 79

Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 140203850X. 43

Panagiotis G. Ipeirotis and Luis Gravano. Classification-aware hidden-web text database selection. *ACM Trans. Inf. Syst.*, 26(2):6:1–6:66, April 2008. ISSN 1046-8188. doi: 10.1145/1344411.1344412. 52

Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern Recogn.*, 38(12):2270–2285, December 2005. ISSN 0031-3203. doi: 10.1016/j.patcog.2005.01.012. 129

Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. ISBN 0-13-022278-X. 57

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Inf. Process. Manage.*, 36 (2):207–227, January 2000a. ISSN 0306-4573. doi: 10.1016/S0306-4573(99)00056-4. 89

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000b. ISSN 0306-4573. doi: 10.1016/S0306-4573(99)00056-4. 20

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. volume 20, pages 422–446. ACM, 2002. 45

Hai Jin, Xiaomin Ning, and Hanhua Chen. Efficient search for peer-to-peer information retrieval using semantic small world. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 1003–1004, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135986. 74

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775067. 47, 62

Hideo Joho, Leif A. Azzopardi, and Wim Vanderbauwhede. A survey of patent users: An analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, pages 13–24, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0247-0. doi: 10.1145/1840784.1840789. 21

K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: Development and comparative experiments. *Inf. Process. Manage.*, 36(6): 779–808, November 2000. ISSN 0306-4573. doi: 10.1016/S0306-4573(00)00015-7. 90

Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43(2):618–644, March 2007. ISSN 0167-9236. doi: 10.1016/j.dss.2005.05.019. 9, 12, 78, 136, 137, 138

Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 640–651, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. doi: 10.1145/775152.775242. 78

Gabriella Kazai and Natasa Milic-Frayling. Trust, authority and popularity in social information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1503–1504, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458356. 79, 137

Weimao Ke and Javed Mostafa. Studying the clustering paradox and scalability of search in highly distributed environments. *ACM Trans. Inf. Syst.*, 31(2):8:1–8:36, May 2013. ISSN 1046-8188. doi: 10.1145/2457465.2457468. 77

Diane Kelly and Nicholas J. Belkin. Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 408–409, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. doi: 10.1145/383952.384045. 47

Allen Kent, M. M. Berry, and J. W. Perry. Machine literature searching ii. problems in indexing for machine searching. *American Documentation*, 5(1):22–25, 1954. ISSN 1936-6108. 44

Youngho Kim and W. Bruce Croft. Diversifying query suggestions based on query documents. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 891–894,

New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. doi: 10.1145/2600428.2609467. 21

Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 193–202, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2351-2. doi: 10.1145/2556195.2556220. 47

Iraklis A. Klampanos and Joemon M. Jose. An architecture for peer-to-peer information retrieval. In *Proc. of SIGIR*, pages 401–402, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860521. 6, 74, 84, 92

Iraklis A Klampanos and Joemon M Jose. An architecture for information retrieval over semi-collaborating peer-to-peer networks. In *In Proceedings of the 19th ACM Symposium on Applied Computing (SAC 2004*, pages 1078–1083, 2004. 64, 74

Iraklis A. Klampanos and Joemon M. Jose. An evaluation of a cluster-based architecture for peer-to-peer information retrieval. In *DEXA*, pages 380–391, 2007. 5, 6, 7, 8, 64, 69, 70, 71, 72, 74, 84, 92, 93, 96, 101, 161

Iraklis A. Klampanos and Joemon M. Jose. Searching in peer-to-peer networks. *Computer Science Review*, 6(4):161 – 183, 2012. ISSN 1574-0137. doi: 10.1016/j.cosrev.2012.07.001. 4, 6, 51, 53, 63, 70

Iraklis A. Klampanos, Victor Poznański, Joemon M. Jose, Peter Dickman, and Edmund Halley Road. A suite of testbeds for the realistic evaluation of peer-to-peer information retrieval systems. In *Proc. of ECIR*, pages 38–51, 2005. 8, 71, 75, 87, 163, 165

Iraklis A. Klampanos, Joemon M. Jose, and C. J. "Keith" van Rijsbergen. Single-pass clustering for peer-to-peer information retrieval: the effect of document ordering. In *Proceedings of the 1st international conference on Scalable information systems*, InfoScale '06, New York, NY, USA, 2006. ACM. ISBN 1-59593-428-6. doi: 10.1145/1146847.1146883. 75

Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, pages 163–170, New York, NY, USA, 2000. ACM. ISBN 1-58113-184-4. doi: 10.1145/335305.335325. 6, 70, 84

Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1168–1176, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2488217. 43

Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. Aggregated search: A new information retrieval paradigm. *ACM Comput. Surv.*, 46(3):41:1–41:31, January 2014. ISSN 0360-0300. doi: 10.1145/2523817. 4, 51

Timo Koskela, Otso Kassinen, Erkki Harjula, and Mika Ylianttila. P2p group management systems: A conceptual analysis. *ACM Comput. Surv.*, 45(2):20:1–20:25, March 2013. ISSN 0360-0300. doi: 10.1145/2431211.2431219. 70

Balachander Krishnamurthy and Jia Wang. On network-aware clustering of web clients. *SIGCOMM Comput. Commun. Rev.*, 30(4):97–110, August 2000. ISSN 0146-4833. doi: 10.1145/347057.347412. 69

Narayanan Kulathuramaiyer and Wolf-Tilo Balke. Restricting the view and connecting the dots - dangers of a web search engine monopoly. *J. UCS*, 12(12):1731–1740, 2006. 48

Anagha Kulkarni and Jamie Callan. Selective search: Efficient and effective search of large textual collections. *ACM Trans. Inf. Syst.*, 33(4):17:1–17:33, April 2015. ISSN 1046-8188. doi: 10.1145/2738035. 90

Anagha Kulkarni, Almer S. Tigelaar, Djoerd Hiemstra, and Jamie Callan. Shard ranking and cutoff estimation for topically partitioned collections. In *Proceedings of CIKM '12*, CIKM '12, pages 555–564, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2396833. 52

Giridhar Kumaran and Vitor R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 564–571, Boston, MA, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572038. 19, 20

John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 111–119, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. doi: 10.1145/383952.383970. 31, 32, 57

Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, New Orleans, LA, USA, 2001. ACM. ISBN 1-58113-331-6. doi: http://doi.acm.org/10.1145/383952.383972. 32

Joon Ho Lee. Analyses of multiple evidence combination. *SIGIR Forum*, 31(SI):267–276, July 1997. ISSN 0163-5840. doi: 10.1145/278459.258587. 91

Dirk Lewandowski. The retrieval effectiveness of web search engines: Considering results descriptions. *CoRR*, abs/1511.05800, 2015. 89

Dirk Lewandowski, Henry Wahlig, and Gunnar Meyer-Bautor. The freshness of web search engine databases. *J. Inf. Sci.*, 32(2):131–148, April 2006. ISSN 0165-5515. doi: 10.1177/0165551506062326. 3, 48

Hanbing Li. Exploring the Popularity, Reputation and Certification of User-Generated Software. Master's thesis, IRISA-Télécom Bretagne Brest, équipe REOP, September 2012. 139

Hang Li. Query understanding in web search: By large scale log data mining and statistical learning. In *Proceedings of the 2nd Workshop on NLP Challenges in the Information Explosion Era*, page 1, Beijing, China, 2010. 19

Hang Li and Jun Xu. Machine learning for query-document matching in search. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 767–768. ACM, 2012. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124393. 21

Juan Li and Son Vuong. Soon: A scalable self-organized overlay network for distributed information retrieval. In *Proceedings of the 19th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management: Managing Large-Scale Service Deployment*, DSOM '08, pages 1–13, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85999-4. doi: 10.1007/978-3-540-87353-2_1. 75

Mu Li, Muhua Zhu, Yang Zhang, and Ming Zhou. Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics*, pages 1025–1032, Sydney, Australia, 2006. ACL. 19, 20

P. Li, C.J.C. Burges, and Q. Wu. Learning to rank using classification and gradient boosting. In *Advances in Neural Information Processing Systems 20*, number MSR-TR-2007-74, page 0. MIT Press, Cambridge, MA, January 2008. 38

Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3): 225–331, March 2009. ISSN 1554-0669. doi: 10.1561/1500000016. 39

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011. ISBN 978-3-642-14266-6. xi, 9, 37, 38, 39

Jie Lu. Full-text federated search in peer-to-peer networks. *SIGIR Forum*, 41(1):121–121, June 2007. ISSN 0163-5840. doi: 10.1145/1273221.1273233. 4, 5, 6, 7, 8, 53, 63, 68, 69, 70, 72, 74

Jie Lu and Jamie Callan. Content-based retrieval in hybrid peer-to-peer networks. In *Proc.of CIKM*, pages 199–206, 2003. 6, 64, 73, 165

Jie Lu and Jamie Callan. Federated search of text-based digital libraries in hierarchical peer-to-peer networks. In *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, pages 52–66, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-25295-9, 978-3-540-25295-5. doi: 10.1007/978-3-540-31865-1_5. 74

Jie Lu and Jamie Callan. User modeling for full-text federated search in peer-to-peer networks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 332–339, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148229. 74

Eng Keong Lua, Jon Crowcroft, Marcelo Pias, Ravi Sharma, and Steven Lim. A survey and comparison of peer-to-peer overlay network schemes. *IEEE Communications Surveys and Tutorials*, 7:72–93, 2005. 63

H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957. ISSN 0018-8646. doi: 10.1147/rd.14.0309. 3, 19, 22

Qin Lv, Pei Cao, Edith Cohen, Kai Li, and Scott Shenker. Search and replication in unstructured peer-to-peer networks. In *Proceedings of the 16th international conference on Supercomputing*, ICS '02, pages 84–95, New York, NY, USA, 2002. ACM. ISBN 1-58113-483-5. doi: 10.1145/514191.514206. 65

Craig Macdonald, Rodrygo L.T. Santos, and Iadh Ounis. On the usefulness of query features for learning to rank. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2559–2562, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398691. 21

D. J. C. Mackay and L. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1994. 34

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999. ISBN 0-262-13360-1. 32

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008a. 15, 18, 21, 33

Christopher D. Manning, Prabhakar Raghavan, and Hinrich SchÃijtze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008b. ISBN 978-0-521-86571-5. 47

A. A. Markov. *Theory of Algorithms*. Academy of Sciences of the USSR, 1954. 29

Ilya Markov and Fabio Crestani. Theoretical, qualitative, and quantitative analyses of small-document approaches to resource selection. *ACM Trans. Inf. Syst.*, 32(2): 9:1–9:37, April 2014. ISSN 1046-8188. doi: 10.1145/2590975. 52, 71, 110

D. Maxwell and L. Azzopardi. Simulating interactive information retrieval. In *Proc. 39th ACM SIGIR*, pages 1141–1144, 2016. 47

D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. Searching and stopping: An analysis of stopping rules and strategies. In *Proc. 24th ACM CIKM*, pages 313–322, 2015. 47, 48

Petar Maymounkov and David Mazières. Kademlia: A peer-to-peer information system based on the xor metric. In *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, IPTPS '01, pages 53–65, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44179-4. 68

Massimo Melucci and Alberto Poggiani. A study of a weighting scheme for information retrieval in hierarchical peer-to-peer networks. In *Proceedings of ECIR*, pages 136–147, 2007. doi: 10.1007/978-3-540-71496-5_15. 52, 59, 122

Weiyi Meng, Clement Yu, and King-Lup Liu. Building efficient and effective metasearch engines. *ACM Comput. Surv.*, 34(1):48–89, March 2002. ISSN 0360-0300. doi: 10.1145/505282.505284. 4, 51, 110

Donald Metzler and W. Bruce Croft. Linear feature-based models for information retrieval. *Inf. Retr.*, 10(3):257–274, June 2007. ISSN 1386-4564. doi: 10.1007/s10791-006-9019-z. 38, 40

Abbe Mowshowitz and Akira Kawaguchi. Assessing bias in search engines. *Inf. Process. Manage.*, 38(1):141–156, 2002. 48

Linh Thai Nguyen, D. Jia, Wai Gen Yee, and Ophir Frieder. Analysis of query logs in gnutella peer-to-peer network. In *Proceedings of the ACM Thirtieth Conference on Research and Development in Information Retrieval*, 2007. 2, 89

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006. 90, 103

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120. 79

Georgios Paltoglou, Michail Salampasis, and Maria Satratzemi. Integral based source selection for uncooperative distributed information retrieval environments. In *Proceedings of LSDS-IR*, pages 67–74, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-254-2. doi: 10.1145/1458469.1458475. 52

Gautam Pant, Padmini Srinivasan, and Filippo Menczer. Crawling the Web. In Mark Levene and Alexandra Poulovassilis, editors, *Web dynamics: Adapting to change in content, size, topology and use*. Springer, 2004. 17

Jarutas Pattanaphanchai, Kieron O'Hara, and Wendy Hall. Trustworthiness criteria for supporting users to assess the credibility of web information. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 1123–1130, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2038-2. 79

Fuchun Peng, Nawaaz Ahmed, Xin Li, and Yumao Lu. Context sensitive stemming for web search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 639–646, Amsterdam, The Netherlands, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277851. 3, 19, 20

José R. Pérez-Agüera and Lourdes Araujo. Comparing and combining methods for automatic query expansion. *CoRR*, abs/0804.2057, 2008. 145

Nina Phan, Peter Bailey, and Ross Wilkinson. Understanding the relationship of information need specificity to search query length. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 709–710, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277870. 46

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998a. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291008. 59

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, Melbourne, Australia, 1998b. ACM. ISBN 1-58113-015-5. 29, 31

Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. 3, 19, 20

Maroje Puh, Toan Luu, Ivana Podnar Zarko, and Martin Rajman. Scalable content-based ranking in p2p information retrieval. In *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part I*, KES '08, pages 633–640, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85562-0. doi: 10.1007/978-3-540-85563-7_80. 71

Diego Puppin, Fabrizio Silvestri, Raffaele Perego, and Ricardo Baeza-Yates. Tuning the capacity of search engines: Load-driven routing and incremental caching to reduce and balance the load. *ACM Trans on Info Syst (TOIS)*, 28(2):5, 2010. 60

Paraskevi Raftopoulou and Euripides G. M. Petrakis. icluster: a self-organizing overlay network for p2p information retrieval. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 65–76, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 3-540-78645-7, 978-3-540-78645-0. 75

Paraskevi Raftopoulou, Euripides G. M. Petrakis, Christos Tryfonopoulos, and Gerhard Weikum. Information retrieval and filtering over self-organising digital libraries. In *ECDL*, pages 320–333, 2008. 75

Vijay V. Raghavan, Gwang S. Jung, and Peter Bollmann. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229, 1989. 44

M. Krishna Ramanathan, V. Kalogeraki, and J. Pruyne. Finding good peers in peer-to-peer networks. In *Parallel and Distributed Processing Symposium., Proceedings International, IPDPS 2002, Abstracts and CD-ROM*, pages 8 pp–, April 2002. doi: 10.1109/IPDPS.2002.1015499. 69

Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. *SIGCOMM Comput. Commun. Rev.*, 31(4): 161–172, August 2001. ISSN 0146-4833. doi: 10.1145/964723.383072. 68

Sami Richardson and Ingemar J. Cox. Estimating global statistics for unstructured p2p search in the presence of adversarial peers. In *Proc. of SIGIR*, pages 203–212, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. doi: 10.1145/2600428.2609567. 5, 8, 54, 71

Ajitha Robert and S. Sendhilkumar. Provenance based web search. In Ajith Abraham and Sabu M Thampi, editors, *Intelligent Informatics*, volume 182 of *Advances in Intelligent Systems and Computing*, pages 451–458. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-32062-0. doi: 10.1007/978-3-642-32063-7_48. 78, 79

S. E. Robertson and K. Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976. 26

S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. 27, 55

S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 35–56. Butterworth & Co., 1981. ISBN 0-408-10775-8. 27

Stephen Robertson. On the optimisation of evaluation metrics. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, Singapore, Singapore, 2008. ACM. 44

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. ISSN 1554-0669. doi: 10.1561/1500000019. 25, 28

Stephen E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977. doi: 10.1108/eb026647. 25

Stephen E. Robertson and Micheline Hancock-Beaulieu. On the evaluation of ir systems. *Inf. Process. Manage.*, 28(4):457–466, 1992. 43

Stephen E. Robertson and Steve Walker. Okapi/Keenbow at TREC-8. In *Proceedings of TREC*, volume 8, 1999. 59

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-2. In *Proceedings of the 2nd Text REtrieval Conference*, Gaithersburg, MD, USA, 1993. 28

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, Gaithersburg, MD, USA, 1994. 28

J. J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971. 20

Dakota S. Rudesill, Caverlee James, and Sui Daniel. The deep web and the darknet: A look inside the internet's massive black box. *Woodrow Wilson International Center for Scholars, STIP 03*, 314, oct 2015. 2, 18, 48

I. Rudomilov and I. Jelínek. Class-based approach in semantic p2p information retrieval. In *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*, pages 257–261, Sept 2012. 76

Sini Ruohomaa and Lea Kutvonen. Trust management survey. In *Proceedings of the Third international conference on Trust Management*, iTrust'05, pages 77–92, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-26042-0, 978-3-540-26042-4. doi: 10.1007/11429760_6. 138

Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, June 2003. ISSN 0269-8889. doi: 10.1017/S0269888903000638. 47

G. Salton, editor. *The SMART Retrieval System Âη Experiments in Automatic Document Processing.* Englewood Cliffs: Prentice-Hall, 1971. 44

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. 3, 19, 23, 87

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0. 22

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840. 3, 18, 55

Gerard Salton, Edward A. Fox, and Ellen M. Voorhees. A comparison of two methods for boolean query relevance feedback. Technical report, Ithaca, NY, USA, 1983. 55

Stefan Saroiu, Krishna P. Gummadi, and Steven D. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *Multimedia Computing and Networking (MMCN)*, January 2002. 137, 140

Jangwon Seo and W. Bruce Croft. Blog site search using resource selection. In *Proceedings of CIKM*, pages 1053–1062, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458222. 52, 61, 62

Yilei Shao and Randolph Wang. Buddynet: History-based p2p search. In *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, pages 23–37, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-25295-9, 978-3-540-25295-5. doi: 10.1007/978-3-540-31865-1_3. 69

Joseph A. Shaw and Edward A. Fox. Combination of Multiple Searches. In *Text REtrieval Conference*, pages 243–252, 1994. 91

Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 131–138, Seattle, WA, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148196. 20

Ji-Rong Wen Wei-Ying Ma Shipeng Yu, Deng Cai. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. Technical report, December 2002. 47

Milad Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. In *Proceedings of ECIR*, pages 160–172, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-71494-1. 52

Milad Shokouhi and Luo Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011. 4, 50, 54, 110

Luo Si and Jamie Callan. Relevant document distribution estimation method for resource selection. In *Proc. of SIGIR*, pages 298–305, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860490. 52, 53

Luo Si and Jamie Callan. Unified utility maximization framework for resource selection. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 32–41, New York, NY, USA, 2004. ACM. ISBN 1-58113-874-1. doi: 10.1145/1031171.1031180. 52

Luo Si, Rong Jin, Jamie Callan, and Paul Ogilvie. A language modeling framework for resource selection and results merging. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 391–397, New York, NY, USA, 2002. ACM. ISBN 1-58113-492-4. doi: 10.1145/584792.584856. 57, 58

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. ISSN 0163-5840. doi: 10.1145/331403.331405. 20

Daria Sorokina, Rich Caruana, and Mirek Riedewald. *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings*, chapter Additive Groves of Regression Trees, pages 323–334. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74958-5. doi: 10.1007/978-3-540-74958-5_31. 38

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. 22

Kunwadee Sripanidkulchai and Hui Zhang. *Web Content Delivery*, chapter Content Location in Peer-to-Peer Systems: Exploiting Locality, pages 73–97. Springer US, Boston, MA, 2005. ISBN 978-0-387-27727-1. doi: 10.1007/0-387-27727-7_4. 69

M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. Technical Report 00-034, University of Minnesota, 2000. 94, 174

Ion Stoica, Robert Morris, David Liben-nowell, David R. Karger, M. Frans Kaashoek, Frank Dabek, and Hari Balakrishnan. Chord: a scalable peer-to-peer lookup protocol for internet applications. *IEEE/ACM Transactions on Networking*, 11:17–32, 2003. 58, 68

Daniel Stutzbach and Reza Rejaie. Characterizing the two-tier gnutella topology. *SIGMETRICS Perform. Eval. Rev.*, 33(1):402–403, June 2005. ISSN 0163-5999. doi: 10.1145/1071690.1064275. 70

Daniel Stutzbach and Reza Rejaie. Understanding churn in peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, IMC '06, pages 189–202, New York, NY, USA, 2006. ACM. ISBN 1-59593-561-4. doi: 10.1145/1177080.1177105. 8, 106, 165

G. M. Tallis. The use of a generalized multinomial distribution in the estimation of correlation in discrete data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):530–534, 1962. ISSN 00359246. 30

Tao Tao and ChengXiang Zhai. A mixture clustering model for pseudo feedback in information retrieval. In David Banks, FrederickR. McMorris, Phipps Arabie, and Wolfgang Gaul, editors, *Classification, Clustering, and Data Mining Applications*, Studies in Classification, Data Analysis, and Knowledge Organisation, pages 541–551. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-22014-5. doi: 10.1007/978-3-642-17103-1_51. 59

Michael Taylor, Hugo Zaragoza, Nick Craswell, Stephen Robertson, and Chris Burges. Optimisation methods for ranking functions with multiple parameters. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 585–593, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. doi: 10.1145/1183614.1183698. 37

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Characterizing the value of personalizing search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 757–758, Amsterdam, The Netherlands, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277894. 45

Omer Tene. What google knows: Privacy and internet search engines. *Utah Law Review*, 2008(4):1434–1490, 2009. 3, 48

Henri Theil. A Rank-Invariant Method of Linear and Polynomial Regression Analysis. In Baldev Raj and Johan Koerts, editors, *Henri Theil's Contributions to Economics and Econometrics*, volume 23 of *Advanced Studies in Theoretical and Applied Econometrics*, pages 345–381. Springer Netherlands, 1992. doi: 10.1007/978-94-011-2546-8_20. 127

Paul Thomas and Milad Shokouhi. Sushi: Scoring scaled samples for server selection. In *Proceedings of SIGIRl*, pages 419–426, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572014. 52

Jing Tian and Yafei Dai. Understanding the dynamic of peer-to-peer systems. In *IPTPS*, 2007. 8

Almer S. Tigelaar, Djoerd Hiemstra, and Dolf Trieschnigg. Peer-to-peer information retrieval: An overview. *ACM Trans. Inf. Syst.*, 30(2):9:1–9:34, May 2012. ISSN 1046-8188. doi: 10.1145/2180868.2180871. 4, 6, 51, 53, 63, 64, 70, 81, 85

H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, pages 1–24, New York, NY, USA, 1990. ACM. ISBN 0-89791-408-2. doi: 10.1145/96749.98006. 55

Howard Robert Turtle. *Inference networks for document retrieval*. PhD thesis, University of Massachusetts, 1991. 55

Julián Urbano, Mónica Marrero, and Diego Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proceedings of SIGIR '13*, pages 925–928, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484163. 151

van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294. 5, 42, 44, 46, 47, 69, 84

Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *J. Mach. Learn. Res.*, 17(1):3581–3618, January 2016. ISSN 1532-4435. 148

Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, 2005. ISBN 0262220733. 41, 42

Wei Wang, Guosun Zeng, and Lulai Yuan. Ant-based reputation evidence distribution in p2p networks. In *Proceedings of the Fifth International Conference on Grid and Cooperative Computing*, GCC '06, pages 129–132, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2694-2. doi: 10.1109/GCC.2006.29. 78

Yao Wang and Julita Vassileva. Bayesian network trust model in peer-to-peer networks. In *Proceedings of the Second international conference on Agents and Peer-to-Peer Computing*, AP2PC'03, pages 23–34, Berlin, Heidelberg, 2004. Springer-Verlag. ISBN 3-540-24053-5, 978-3-540-24053-2. doi: 10.1007/978-3-540-25840-7_3. 78

D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. 6, 70, 84

William Webber and Laurence A. F. Park. Score adjustment for correction of pooling bias. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research*

and Development in Information Retrieval, SIGIR '09, pages 444–451, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572018. 42

Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148204. 122

Colin Wilkie and Leif Azzopardi. *Best and Fairest: An Empirical Analysis of Retrieval System Bias*, pages 13–25. Springer International Publishing, Cham, 2014. ISBN 978-3-319-06028-6. doi: 10.1007/978-3-319-06028-6_2. 145

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0123748569, 9780123748560. 148

Qiang Wu, Christopher J.C. Burges, Krysta Svore, and Jianfeng Gao. Ranking, boosting, and model adaptation. Technical Report MSR-TR-2008-109, Microsoft Research, October 2008. 38

Qianli Xing, Yiqun Liu, Jian-Yun Nie, Min Zhang, Shaoping Ma, and Kuo Zhang. Incorporating user preferences into click models. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1301–1310, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. doi: 10.1145/2505515.2505704. 150

Li Xiong and Ling Liu. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 16:843–857, 2004. 78

Jingfang Xu and Xing Li. Learning to rank collections. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 765–766, New York, NY, USA, 2007a. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277898. 52, 62

Jinxi Xu and James P. Callan. Effective retrieval with distributed collections. In *Proc. of SIGIR*, pages 112–120, 1998. 98

Jinxi Xu and W. Bruce Croft. Cluster-based language models for distributed retrieval. In *Proc. of SIGIR*, pages 254–261, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: 10.1145/312624.312687. 8, 52, 57, 58, 73, 118, 121

Jun Xu and Hang Li. Adarank: A boosting algorithm for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 391–398, New York, NY, USA, 2007b. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277809. 38, 40

Beverly Yang and Hector Garcia-Molina. Improving search in peer-to-peer networks. In *Proceedings of the 22 nd International Conference on Distributed Computing Systems (ICDCS'02)*, ICDCS '02, pages 5–14, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1585-1. 67

Beverly Yang and Hector Garcia-Molina. Designing a super-peer network. In *ICDE*, pages 49–60, 2003. 67, 72, 85

Mengkun Yang and Zongming Fei. A novel approach to improving search efficiency in unstructured peer-to-peer networks. *Journal of Parallel and Distributed Computing*, 69(11):877 – 884, 2009. ISSN 0743-7315. doi: 10.1016/j.jpdc.2009.07.004. 64

T. Yeferny, K. Arour, and A. Bouzeghoub. An efficient peer-to-peer semantic overlay network for learning query routing. In *Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on*, pages 1025–1032, March 2013. doi: 10.1109/AINA.2013.129. 69

Budi Yuwono and Dik Lun Lee. Wise: A world wide web resource database system. *IEEE Trans. on Knowl. and Data Eng.*, 8(4):548–554, August 1996. ISSN 1041-4347. doi: 10.1109/69.536248. 56

Budi Yuwono and Dik Lun Lee. Server ranking for distributed text retrieval systems on the internet. In *DASFAA*, pages 41–50. World Scientific Press, 1997. ISBN 981-02-3107-5. 52, 56, 118, 121, 128

ChengXiang Zhai. Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2008. ISSN 1554-0669. 29, 32, 34

Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410, New York, NY, USA, 2001. ACM. ISBN 1-58113-436-3. doi: 10.1145/502585.502654. 59

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2): 179–214, 2004. 33

Ye Zhang. *Trust-aware information retrieval in peer-to-peer environments*. Phd, 2011. 9, 71, 76, 81, 136, 177, 178

Le Zhao and Jamie Callan. Automatic term mismatch diagnosis for selective query expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 515–524, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348354. 21

Part V

Appendices

In these appendices, I provide supplementary results from research presented in this thesis. In particular, The query routing and resource selection methods were conducted under specific threshold of percentages varies from 5%, and 10% to 50% step 10. In each chapter, I selected one threshold for explanation while the results of other thresholds are exhibited here.

Appendix A

IPI, Document Retrieval and LTRo Approaches

A.1 IPI Retrieval Effectiveness

Table A.1: IPI Retrieval effectiveness and efficiency on Digital Library (∘ & • indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t-test).

DL*		**DLWOR Testbed**							**DLWR Testbed**						
%	Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
100%	Flooding	1174	0.02866	0.54790	0.16900	0.13233	0.08770	0.08659	1187	0.02089	0.42564	0.01837	0.01837	0.02306	0.02232
5%	IPI	**98**	**0.02232**	**0.40398**	**0.17900**•	**0.12833**	**0.07930**	**0.08666**	**90**	**0.02170**	**0.38694**	**0.02700**	**0.02733**	**0.03640**	**0.02625**
	CVV	**88**	0.02203	0.38146	0.16600	0.12232	0.07540	0.07328	**80**	0.02074	0.35934	0.02500	0.02566	0.03640	0.02391
	Taily	**96**	**0.02565**	**0.43992**	**0.17100**	**0.13166**	**0.08370**	**0.09058**	91	**0.02412**	**0.41914**	**0.02800**	0.02667	**0.03710**	**0.02808**
	CORI	**82**	**0.02388**	0.38477	0.17100	0.12400	0.07890	0.07877	**78**	**0.02336**	0.38408	**0.04200**	**0.05600**	**0.05580**	**0.03383**
	KL	105	0.00933	0.07907	0.09400	0.06066	0.02740	0.02037	97	0.00843	0.05678	**0.04600**	**0.03833**	0.02160	0.00811
	vGIOSS	**68**	0.01562	0.22783	0.14900	0.10400	0.05470	0.04797	**76**	0.01499	0.23921	**0.06000**	**0.06133**	**0.04300**	**0.02909**
	RW	166	0.01195	0.12279	0.11200	0.06734	0.03200	0.03754	128	0.01156	0.08101	**0.03600**	**0.03433**	0.02650	0.00821
IPI Rank @ 5%		5	3	2	1	2	2	2	4	3	2	6	5	4	4
20%	IPI	**312**	**0.02686**	**0.52647**	**0.18200**	**0.13533**	**0.08710**	**0.09333**•	**312**	**0.02449**	**0.49688**	**0.02400**	**0.02233**	**0.02570**	**0.02636**
	CVV	**306**	**0.02723**	0.52205	**0.19200**	**0.13667**	**0.08870**	**0.09199**	**294**	0.02263	0.45304	0.01900	0.02000	**0.02650**	0.02466
	Taily	337	**0.02849**	**0.53061**	0.16300	0.13000	**0.08960**	0.09006	332	**0.02580**	**0.50380**	0.02300	0.02133	**0.02710**	**0.02872**
	CORI	**304**	**0.02902**	0.50951	0.17300	**0.14267**	**0.09090**	0.08484	**294**	**0.02495**	0.47197	0.02100	**0.02233**	**0.03040**	**0.02795**
	KL	395	0.01620	0.24897	0.13000	0.09233	0.05850	0.04220	373	0.01473	0.23504	0.02400	**0.02933**	**0.03540**	0.01773
	vGIOSS	292	0.02557	0.48414	**0.18600**	**0.13567**	0.08440	0.08489	**292**	0.02249	0.41386	0.02200	**0.02666**	**0.04200**	**0.02835**
	RW	522	0.01930	0.30321	0.14600	0.10533	0.06290	0.04613	437	0.01586	0.22066	**0.02600**	**0.02900**	**0.02760**	0.01466
IPI Rank @ 20%		4	4	2	3	4	4	1	4	3	2	2	4	6	4
30%	IPI	**450**	**0.02817**	**0.55086**•	**0.16400**	**0.13200**	**0.08960**	**0.08896**	**452**	**0.02497**	**0.50091**	**0.02300**	**0.02000**	**0.02320**	**0.02583**
	CVV	**440**	**0.02832**	0.53794	**0.18400**	**0.13366**	0.08840	**0.08953**	**428**	0.02199	0.45441	0.01800	0.01933	**0.02430**	0.02355
	Taily	493	**0.02909**	**0.54828**	0.15700	0.12633	**0.09020**	0.08729	491	**0.02564**	**0.50108**	0.02200	**0.02000**	**0.02510**	**0.02772**
	CORI	**440**	**0.02929**	0.53419	**0.16800**	**0.13700**	**0.09110**	0.08557	**427**	0.02479	0.47308	0.01900	**0.02033**	**0.02680**	**0.02638**
	KL	564	0.01922	0.31722	0.14700	0.09999	0.06590	0.05050	528	0.01772	0.29294	**0.02300**	**0.02300**	**0.03250**	0.02033
	vGOISS	**426**	0.02716	0.52556	**0.17400**	**0.13233**	0.08750	**0.09071**	**427**	0.02379	0.45443	0.02000	**0.02466**	**0.03520**	**0.02779**
	RW	681	0.02166	0.36004	0.13600	0.10766	0.06980	0.06526	595	0.02003	0.31353	0.01900	0.02433	**0.02870**	0.01755
IPI Rank @ 30%		4	4	1	4	4	3	3	4	2	2	1	5	7	4
40%	IPI	**581**	**0.02849**	**0.55689**∘	**0.13600**	**0.12533**	**0.08740**	**0.08499**	**582**	**0.02482**	**0.50014**∘	**0.02200**	**0.02000**	**0.02270**	**0.02566**
	CVV	**568**	**0.02862**	0.54157	**0.16500**	**0.13000**	**0.08930**	0.08391	**555**	0.02153	0.43317	0.01900	0.01833	**0.02330**	0.02288
	Taily	637	**0.02904**	**0.55292**	**0.14600**	0.12267	**0.08840**	**0.08732**	635	**0.02529**	**0.49908**	**0.02200**	**0.02033**	**0.02420**	**0.02673**
	CORI	**568**	**0.02948**	0.55202	**0.16900**	**0.13366**	**0.09040**	**0.08673**	**554**	0.02392	0.47566	0.01800	0.01933	**0.02500**	0.02513
	KL	708	0.02240	0.39749	**0.15300**	0.11233	0.07340	0.06295	665	0.02006	0.36329	0.01900	**0.02133**	**0.02950**	0.02259
	vGOISS	**556**	0.02811	0.53928	**0.16100**	**0.12900**	0.08680	**0.08503**	**555**	0.02428	0.46953	0.01900	**0.02100**	**0.02690**	**0.02627**
	RW	801	0.02367	0.40191	0.13600	0.11867	0.07430	0.05530	726	0.02045	0.32883	0.02000	**0.02300**	**0.02630**	0.02092
IPI Rank @ 40%		4	4	1	6	4	4	4	4	2	1	1	5	7	3
50%	IPI	**709**	**0.02860**	**0.55770**∘	**0.12400**	**0.12000**	**0.08670**	**0.08017**	**714**	**0.02482**	**0.50134**∘	**0.02100**	**0.01967**	**0.02240**	**0.02520**
	CVV	**692**	**0.02889**	0.54909	**0.16500**	**0.13066**	**0.08890**	**0.08371**	**679**	0.02126	0.43806	0.01800	0.01800	**0.02320**	0.02268
	Taily	776	**0.02895**	**0.55308**	**0.14400**	**0.12333**	**0.08780**	**0.08648**	773	**0.02503**	**0.49994**	0.02100	**0.01967**	**0.02380**	**0.02610**
	CORI	**693**	**0.02930**	0.55217	**0.16000**	**0.12966**	**0.08970**	**0.08468**	**676**	0.02324	0.46961	0.01800	0.01900	**0.02380**	0.02434
	KL	842	0.02472	0.44224	**0.16100**	0.11666	0.07840	0.06798	791	0.02152	0.39351	0.01900	**0.02066**	**0.02770**	0.02336
	vGIOSS	**682**	0.02854	0.54106	**0.14500**	**0.12700**	**0.08690**	**0.08155**	**678**	0.02332	0.46587	0.01800	0.01933	**0.02420**	0.02421
	RW	894	0.02419	0.45144	**0.13800**	0.11966	0.07770	0.07167	833	0.01921	0.34501	0.01400	0.02067	**0.02320**	0.01660
IPI Rank @ 50%		4	4	1	7	5	5	5	4	2	1	1	3	7	2

Table A.2: IPI Retrieval effectiveness and efficiency on File sharing ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t-test).

ASIS*		**ASISWOR Testbed**							**ASISWR Testbed**						
%	Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
100%	Flooding	4963	0.02532	0.46296	0.16400	0.11966	0.07620	0.07122	5279	0.01635	0.33847	0.01414	0.01818	0.01899	0.01732
5%	IPI	**431**	**0.02354**	**0.43642**	**0.16000**	**0.11766**	**0.07630**	**0.06864**	**464**	**0.02085**	**0.41596**	**0.01700**	**0.02133**	**0.02280**	**0.02048**
	CVV	**420**	**0.02384**	**0.43915**	**0.16000**	**0.12200**	0.07530	**0.06940**	**443**	**0.02121**	**0.41605**	0.01600	**0.02167**	**0.02390**	**0.02131**
	Taily	548	**0.02577**	**0.44469**	0.15400	**0.12267**	**0.07820**	**0.07047**	565	**0.02209**	0.41179	**0.02000**	**0.02200**	**0.02590**	**0.02338**
	CORI	433	**0.02566**	**0.43851**	**0.16500**	**0.12466**	**0.07660**	**0.07411**	**455**	**0.02316**	**0.42360**	**0.01700**	**0.02333**	**0.02860**	**0.02351**
	KL	593	0.01174	0.18069	0.10900	0.07433	0.04200	0.03286	593	0.01155	0.18331	**0.02100**	**0.03300**	**0.03590**	0.01559
	vGIOSS	488	0.01589	0.29820	0.12300	0.09333	0.05350	0.05675	501	0.01713	0.32008	**0.02000**	**0.03933**	**0.04390**	**0.02694**
	RW	535	0.01185	0.10320	0.07800	0.05400	0.02920	0.01721	570	0.01309	0.10528	**0.02300**	**0.02700**	**0.02310**	0.00855
IPI Rank @ 5%		2	4	4	2	4	3	4	3	4	3	5	7	7	5
20%	IPI	**1534**	**0.02526**	**0.46707**	**0.16300**	**0.11900**	**0.07710**	**0.07113**	**1643**	**0.01943**	**0.38872**	**0.01500**	**0.01800**	**0.01940**	**0.01846**
	CVV	**1499**	**0.02554**	0.46542	**0.16500**	**0.11966**	0.07670	**0.07184**	**1595**	0.01942	**0.38981**	0.01400	**0.01867**	**0.02060**	**0.01864**
	Taily	2636	**0.02608**	0.46309	0.16000	**0.11966**	**0.07830**	0.07104	2640	**0.01946**	0.37962	**0.01900**	**0.01966**	**0.02180**	**0.02040**
	CORI	1553	**0.02632**	**0.47639**	**0.16600**	**0.12266**	**0.07870**	**0.07484**	**1639**	0.01859	0.37742	0.01300	**0.01800**	**0.02060**	0.01830
	KL	2065	0.01854	0.35115	0.13400	0.10066	0.06300	0.05238	2101	0.01786	0.35171	0.01400	**0.01933**	**0.02100**	0.01823
	vGIOSS	1723	0.02357	0.42905	**0.17300**	**0.12233**	0.07690	**0.07549**	1792	**0.02205**	**0.42937**	0.01300	**0.02033**	**0.02370**	**0.02182**
	RW	1803	0.01708	0.24551	0.12900	0.09499	0.05470	0.03257	1918	0.01657	0.24774	**0.01800**	**0.02233**	**0.02050**	0.01336
IPI Rank @ 20%		2	4	2	4	4	3	4	3	3	3	3	6	7	4
30%	IPI	**2197**	**0.02539**	**0.46934**	**0.16300**	**0.11900**	**0.07710**	**0.07125**	**2243**	**0.01891**	**0.37975**	**0.01400**	**0.01800**	**0.01940**	**0.01807**
	CVV	**2147**	**0.02555**	0.46549	**0.16400**	**0.11966**	0.07640	**0.07179**	2271	0.01853	0.37236	**0.01400**	**0.01800**	0.01930	0.01792
	Taily	4152	**0.02574**	0.46509	0.15900	0.11866	**0.07770**	0.07013	4122	0.01850	0.36757	**0.01600**	**0.01867**	**0.02110**	**0.01948**
	CORI	2219	**0.02605**	**0.47195**	**0.16500**	**0.12133**	**0.07810**	**0.07386**	2331	0.01711	0.35078	**0.01400**	0.01733	**0.01960**	0.01736
	KL	2807	0.02100	0.39297	0.14200	0.10666	0.06730	0.05951	2866	0.01889	0.37122	**0.01500**	**0.01933**	0.01920	**0.01833**
	vGOISS	2420	0.02483	0.45125	**0.16900**	**0.12199**	**0.07840**	**0.07903**	2519	**0.02121**	**0.40903**	**0.01400**	**0.01833**	**0.01970**	**0.01971**
	RW	2447	0.02066	0.32899	0.14600	0.10266	0.06210	0.05248	2607	0.01758	0.29430	**0.02100**	**0.02000**	**0.02260**	0.01631
IPI Rank @ 30%		2	4	2	4	4	4	4	1	2	2	4	5	5	4
40%	IPI	**2813**	**0.02544**	**0.47029**$^{\bullet}$	**0.16100**	**0.11900**	**0.07680**	**0.07125**	**2978**	**0.01798**	**0.36829**	**0.01400**	**0.01767**	**0.01870**	**0.01768**
	CVV	**2741**	**0.02558**	0.46588	**0.16300**	**0.11966**	0.07640	**0.07149**	**2892**	0.01741	0.35826	**0.01400**	**0.01800**	**0.01950**	**0.01768**
	Taily	5609	**0.02556**	0.46440	0.15800	**0.11900**	**0.07760**	0.06978	5547	0.01788	0.36117	**0.01500**	**0.01900**	**0.02070**	**0.01876**
	CORI	2815	**0.02587**	**0.46870**	**0.16300**	**0.12033**	**0.07760**	**0.07274**	**2957**	0.01638	0.33771	**0.01400**	0.01733	**0.01910**	0.01699
	KL	3414	0.02274	0.41503	0.14700	0.11033	0.07190	0.06422	3518	**0.01892**	**0.38250**	**0.01400**	**0.01867**	**0.01900**	**0.01810**
	vGOISS	3029	**0.02557**	0.45935	**0.17000**	**0.12266**	**0.07930**	**0.07447**	3156	**0.01996**	**0.39114**	**0.01400**	**0.01767**	**0.01950**	**0.01871**
	RW	2989	0.01976	0.33711	0.13100	0.09900	0.06320	0.04688	3177	**0.01834**	0.32480	**0.02100**	**0.01867**	**0.02270**	0.01631
IPI Rank @ 40%		2	5	1	4	4	4	4	3	4	3	3	5	7	4
50%	IPI	**3393**	**0.02544**	**0.46952**$^{\circ}$	**0.16100**	**0.11900**	**0.07690**	**0.07122**	**3578**	**0.01754**	**0.36386**	**0.01400**	**0.01733**	**0.01870**	**0.01755**
	CVV	**3289**	**0.02549**	0.46473	**0.16300**	**0.11966**	0.07640	**0.07144**	**3465**	0.01749	0.35950	**0.01400**	**0.01767**	**0.01900**	0.01743
	Taily	7041	**0.02556**	**0.46792**	0.16000	**0.11900**	**0.07760**	0.07035	6978	0.01726	0.35735	**0.01400**	**0.01933**	**0.02040**	**0.01832**
	CORI	**3358**	**0.02576**	0.46707	**0.16400**	**0.12033**	**0.07730**	**0.07222**	**3531**	0.01589	0.32790	**0.01400**	**0.01733**	**0.01910**	0.01680
	KL	3893	0.02353	0.43121	0.15500	0.11399	0.07400	0.06642	4047	**0.01890**	**0.37304**	**0.01400**	**0.01800**	**0.01870**	**0.01779**
	vGIOSS	3562	**0.02588**	0.46205	**0.16800**	**0.12333**	**0.07980**	**0.07403**	3724	**0.01907**	**0.38323**	**0.01400**	**0.01767**	**0.01890**	**0.01807**
	RW	3442	0.02211	0.37198	0.13900	0.10600	0.06680	0.05370	3657	0.01703	0.29759	**0.02000**	**0.01800**	**0.02010**	0.01548
IPI Rank @ 50%		3	5	1	4	4	4	4	3	3	3	2	6	6	4

Table A.3: IPI Retrieval effectiveness and efficiency on Uniformly Distributed ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t-test).

U*		**UWOR Testbed**							**UWR Testbed**						
%	Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
100%	Flooding	2499	0.02764	0.49910	0.21200	0.14800	0.09580	0.10581	1337	0.02269	0.42657	0.01400	0.01900	0.02450	0.02331
5%	IPI	**687**	**0.02319**	**0.42920**	**0.21100**$^{\bullet}$	**0.14933**$^{\bullet}$	**0.08660**	**0.09444**$^{\bullet}$	**702**	**0.02164**	**0.39505**	**0.01600**	**0.02333**	**0.02820**	**0.02352**
	CVV	**657**	0.01767	0.30214	0.17500	0.11333	0.06420	0.06646	732	0.01697	0.30835	0.01200	0.01933	0.02270	0.01708
	Taily	847	**0.02589**	0.43444	**0.18000**	**0.13966**	**0.08770**	**0.08986**	860	**0.02470**	**0.42483**	**0.01700**	**0.02500**	**0.03150**	**0.02852**
	CORI	**608**	**0.02395**	0.40440	0.17600	0.13033	0.08540	0.07654	**661**	**0.02434**	**0.42469**	**0.01700**	**0.03433**	**0.04720**	**0.03239**
	KL	841	0.01047	0.13661	0.10200	0.06633	0.03330	0.03441	1037	0.01005	0.16173	**0.01900**	**0.02333**	0.02370	0.01194
	vGIOSS	**605**	0.01490	0.24839	0.13500	0.09066	0.05190	0.04225	**648**	0.01556	0.28677	**0.01800**	**0.03667**	**0.04140**	0.02018
	RW	1117	0.01565	0.08328	0.11500	0.06633	0.03180	0.01959	1327	0.01392	0.11845	**0.02900**	**0.02666**	0.01920	0.00804
IPI Rank @ 5%		4	3	2	1	1	2	1	3	3	3	6	5	4	3
20%	IPI	**2499**	**0.02764**	**0.49910**	**0.21200**$^{\bullet}$	**0.14800**$^{\bullet}$	**0.09580**	**0.10581**$^{\circ}$	**2507**	**0.02269**	**0.43846**	**0.01300**	**0.01700**	**0.02190**	**0.02215**
	CVV	**2363**	0.02520	0.45742	0.19300	**0.14667**	0.09100	0.09745	2563	0.02080	0.39984	**0.01300**	**0.01767**	0.02160	0.02061
	Taily	3447	**0.02862**	**0.50309**	0.19700	0.14499	**0.09710**	**0.10476**	3487	**0.02467**	**0.45741**	**0.01400**	**0.01767**	**0.02580**	**0.02504**
	CORI	**2204**	**0.02914**	**0.50222**	**0.19800**	0.14500	**0.09850**	0.10345	**2364**	**0.02540**	**0.47351**	0.01200	0.01600	**0.02300**	**0.02500**
	KL	3027	0.01727	0.29480	0.15400	0.10866	0.06350	0.06229	3608	0.01731	0.32185	**0.01300**	**0.01900**	**0.02320**	0.01766
	vGIOSS	**2193**	0.02255	0.42168	0.17800	0.12867	0.08170	0.08684	**2326**	0.02203	0.42049	0.01200	**0.02333**	**0.03160**	**0.02493**
	RW	3581	0.01862	0.24733	0.15500	0.10900	0.06080	0.04517	4090	0.01743	0.25501	**0.01900**	**0.02167**	0.02160	0.01265
IPI Rank @ 20%		4	3	3	1	1	3	1	3	3	3	3	6	5	4
30%	IPI	**3565**	**0.02857**	**0.51773**	**0.21400**$^{\bullet}$	**0.14567**$^{\bullet}$	**0.09520**	**0.10703**$^{\bullet}$	**3578**	**0.02201**	**0.43345**	**0.01300**	**0.01633**	**0.02070**	**0.02120**
	CVV	**3371**	0.02704	0.47601	0.19300	**0.14333**	0.09300	0.10035	3641	0.02143	0.41303	**0.01300**	**0.01700**	0.02060	0.02096
	Taily	5139	0.02860	0.51461	0.18800	0.14000	**0.09570**	0.10377	5190	**0.02347**	**0.44768**	**0.01300**	**0.01633**	**0.02380**	**0.02323**
	CORI	**3180**	**0.02960**	**0.51962**	**0.20200**	0.14300	**0.09880**	**0.10423**	**3404**	**0.02445**	**0.47142**	**0.01300**	0.01567	**0.02150**	**0.02367**
	KL	4173	0.02014	0.34941	0.16500	0.11866	0.07420	0.07051	4771	0.01963	0.37313	**0.01300**	**0.01767**	**0.02210**	0.01962
	vGOISS	**3168**	0.02506	0.45716	0.18300	0.13867	0.08860	0.09153	**3361**	**0.02332**	**0.43796**	**0.01400**	**0.01900**	**0.02420**	**0.02414**
	RW	4697	0.02177	0.31995	0.17300	0.11767	0.07030	0.05665	5260	0.01819	0.30229	0.01200	**0.01767**	0.01890	0.01379
IPI Rank @ 30%		4	3	2	1	1	3	1	3	4	4	2	5	5	4
40%	IPI	**4546**	**0.02875**	**0.51995**$^{\circ}$	**0.20500**$^{\circ}$	**0.14333**$^{\bullet}$	**0.09520**$^{\bullet}$	**0.10663**$^{\bullet}$	**4557**	**0.02130**	**0.42329**	**0.01300**	**0.01567**	**0.02010**	**0.02065**
	CVV	**4291**	0.02806	0.49161	**0.20000**	**0.14033**	0.09380	0.09972	4607	0.02123	0.41916	**0.01300**	**0.01700**	**0.02010**	**0.02065**
	Taily	6650	0.02855	**0.51672**	0.18800	0.13700	0.09290	**0.10322**	6676	**0.02229**	**0.42539**	**0.01400**	**0.01600**	**0.02180**	**0.02184**
	CORI	**4109**	**0.02952**	0.51493	0.19200	0.14000	**0.09460**	0.10277	**4381**	**0.02329**	**0.45334**	0.01200	**0.01567**	**0.02070**	**0.02226**
	KL	5226	0.02256	0.39250	0.18700	0.12633	0.07980	0.07919	5805	0.02022	0.38918	**0.01300**	**0.01700**	**0.02070**	0.02000
	vGOISS	**4094**	0.02640	0.47397	0.19000	0.13833	0.09110	0.09232	**4334**	**0.02300**	**0.43966**	**0.01400**	**0.01767**	**0.02190**	**0.02290**
	RW	5545	0.02227	0.37390	0.17500	0.11967	0.07550	0.06813	6103	0.01956	0.31970	**0.01500**	**0.02033**	**0.02030**	0.01650
IPI Rank @ 40%		4	2	1	1	1	1	1	3	4	4	3	6	6	4
50%	IPI	**5442**	**0.02893**	**0.52064**$^{\bullet}$	**0.20300**	**0.14333**$^{\circ}$	**0.09480**	**0.10679**$^{\bullet}$	**5460**	**0.02068**	**0.41528**	**0.01200**	**0.01567**	**0.02000**	**0.02021**
	CVV	**5144**	0.02854	0.49738	**0.20300**	**0.14033**	0.09440	0.09896	5496	**0.02095**	**0.41891**	**0.01200**	**0.01633**	**0.02000**	**0.02039**
	Taily	8062	0.02833	**0.51341**	0.19100	0.13800	0.09220	**0.10280**	8047	**0.02165**	**0.42222**	**0.01200**	**0.01600**	**0.02060**	**0.02095**
	CORI	**4981**	**0.02940**	0.50998	0.19700	0.13833	**0.09440**	0.10172	**5305**	**0.02213**	**0.43374**	**0.01200**	**0.01567**	**0.02010**	**0.02129**
	KL	6144	0.02490	0.42939	0.18300	0.13300	0.08550	0.08557	6690	**0.02126**	0.41381	**0.01200**	**0.01600**	**0.02020**	**0.02055**
	vGIOSS	**4969**	0.02743	0.48502	0.19700	0.13533	0.09000	0.09404	**5258**	**0.02222**	**0.43483**	**0.01300**	**0.01567**	**0.02060**	**0.02135**
	RW	6191	0.02438	0.38815	0.18600	0.12667	0.08210	0.06903	6732	0.01956	0.34757	**0.01400**	**0.01833**	0.01990	0.01710
IPI Rank @ 50%		4	2	1	1	1	1	1	3	6	5	3	5	5	6

A.2 Documents Retrieval Resource Selection Effectiveness

Table A.4: Document Retrieval effectiveness and efficiency on DL* Testbeds ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t test).

DL*		DLWOR Testbed							DLWR Testbed						
%	Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
100%	Flooding	1174	0.02866	0.54790	0.16900	0.13233	0.08770	0.08659	1187	0.02089	0.42564	0.01837	0.01837	0.02306	0.02232
5%	CVV	88	0.02203	0.38146	0.16600	0.12232	0.07540	0.07328	80	0.02074	0.35934	0.02500	0.02566	0.03640	0.02391
	Taily	96	**0.02565**	**0.43992**	**0.17100**	**0.13166**	**0.08370**	**0.09058**	91	**0.02412**	**0.41914**	0.02800	0.02667	0.03710	0.02808
	CORI	82	0.02388	0.38477	**0.17100**	0.12400	0.07890	0.07877	78	0.02336	0.38408	0.04200	0.05600	**0.05580**	**0.03383**
	KL	105	0.00933	0.07907	0.09400	0.06066	0.02740	0.02037	97	0.00843	0.05678	0.04600	0.03833	0.02160	0.00811
	vGIOSS	**68**	0.01562	0.22783	0.14900	0.10400	0.05470	0.04797	**76**	0.01499	0.23921	**0.06000**	**0.06133**	0.04300	0.02909
	RW	166	0.01195	0.12279	0.11200	0.06734	0.03200	0.03754	128	0.01156	0.08101	0.03600	0.03433	0.02650	0.00821
	BB2	83	**0.02680**$^{\circ}$	0.43427	**0.19800**$^{\bullet}$	**0.14700**$^{\bullet}$	**0.09160**$^{\bullet}$	**0.09819**$^{\bullet}$	80	**0.02603**$^{\bullet}$	0.42672	0.04400	0.05300	**0.05850**$^{\bullet}$	**0.04164**$^{\bullet}$
	In_expB2	83	0.02559	0.41538	**0.20100**$^{\bullet}$	**0.14433**$^{\circ}$	**0.08780**$^{\bullet}$	0.08837	79	**0.02541**	0.40496	0.04500	0.05533	**0.05720**	**0.03825**$^{\bullet}$
	In_expC2	83	0.02551	0.41065	**0.19900**$^{\bullet}$	**0.14466**$^{\bullet}$	**0.08720**$^{\bullet}$	0.08687	79	**0.02520**$^{\circ}$	0.40319	0.04700	0.05600	**0.05820**$^{\circ}$	**0.03833**$^{\bullet}$
	InL2	82	**0.02694**$^{\circ}$	0.42361	**0.20600**$^{\bullet}$	**0.14166**	**0.09230**$^{\bullet}$	**0.09168**$^{\circ}$	79	**0.02683**$^{\bullet}$	0.41384	0.04700	**0.06433**$^{\bullet}$	**0.06540**$^{\bullet}$	**0.04184**$^{\bullet}$
	Hiemstra_LM	83	0.02379	0.36514	**0.19000**$^{\bullet}$	**0.13433**	0.08300	0.08272	78	0.02336	0.35555	**0.08700**$^{\bullet}$	**0.09800**$^{\bullet}$	**0.07190**$^{\bullet}$	**0.04837**$^{\bullet}$
	DFI0	81	0.02557	0.41548	**0.19100**$^{\bullet}$	**0.13633**	**0.08780**$^{\bullet}$	0.08716	77	0.02494	0.40459	0.05700	**0.06933**$^{\bullet}$	**0.06710**$^{\bullet}$	**0.04237**$^{\bullet}$
	TF_IDF	82	**0.02759**$^{\bullet}$	0.43534	**0.21100**$^{\bullet}$	**0.14933**$^{\bullet}$	**0.09540**$^{\bullet}$	**0.09727**$^{\bullet}$	78	**0.02735**$^{\bullet}$	**0.42783**$^{\circ}$	0.04600	**0.06400**$^{\circ}$	**0.06540**$^{\bullet}$	**0.04189**$^{\bullet}$
	BM25	**64**	0.01380	0.16664	0.08100	0.06067	0.04080	0.04047	**74**	0.01451	0.22681	0.04100	0.03400	0.03180	0.02528
Majority @ 5%		1	3	-	7	6	7	3	1	6	2	1	4	7	7
20%	CVV	306	0.02723	0.52205	**0.19200**	0.13667	0.08870	**0.09199**	294	0.02263	0.45304	0.01900	0.02000	0.02650	0.02466
	Taily	337	0.02849	**0.53061**	0.16300	0.13000	0.08960	0.09006	332	**0.02580**	**0.50380**	0.02300	0.02133	0.02710	**0.02872**
	CORI	304	**0.02902**	0.50951	0.17300	**0.14267**	**0.09090**	0.08484	294	0.02495	0.47197	0.02100	0.02233	0.03040	0.02795
	KL	395	0.01620	0.24897	0.13000	0.09233	0.05850	0.04220	373	0.01473	0.23504	0.02400	**0.02933**	0.03540	0.01773
	vGIOSS	**292**	0.02557	0.48414	0.18600	0.13567	0.08440	0.08489	**292**	0.02249	0.41386	0.02200	0.02666	**0.04200**	0.02835
	RW	522	0.01930	0.30321	0.14600	0.10533	0.06290	0.04613	437	0.01586	0.22066	**0.02600**	0.02900	0.02760	0.01466
	BB2	305	**0.02945**	0.53053	0.18700	**0.14466**$^{\circ}$	**0.09480**$^{\bullet}$	**0.09552**$^{\bullet}$	295	0.02534	0.49116	0.02100	0.02100	0.03080	0.02854
	In_expB2	306	**0.02903**	0.52426	0.18100	**0.14466**$^{\circ}$	**0.09330**$^{\bullet}$	0.09188	295	0.02506	0.47775	0.02000	0.02200	0.03030	0.02780
	In_expC2	306	**0.02909**	0.52445	0.17700	**0.14433**	**0.09320**$^{\circ}$	**0.09727**$^{\bullet}$	294	0.02504	0.48706	0.02000	0.02233	0.03070	0.02783
	InL2	304	**0.02967**	0.52186	0.17900	0.14133	**0.09340**$^{\bullet}$	0.08870	293	**0.02609**$^{\bullet}$	0.48400	0.02100	0.02233	0.03170	**0.02934**$^{\circ}$
	Hiemstra_LM	304	0.02881	0.49870	0.18100	0.14100	**0.09380**$^{\bullet}$	0.09084	293	**0.02786**$^{\bullet}$	0.48213	**0.03000**$^{\bullet}$	0.02766	**0.04710**$^{\bullet}$	**0.03599**$^{\bullet}$
	DFI0	298	0.02872	0.51524	**0.19000**	**0.15233**$^{\bullet}$	**0.09570**$^{\bullet}$	**0.10427**$^{\bullet}$	**287**	**0.02727**$^{\bullet}$	0.49313	0.02200	**0.03033**$^{\circ}$	**0.04320**$^{\circ}$	**0.03477**$^{\bullet}$
	TF_IDF	303	**0.02985**$^{\circ}$	**0.53697**$^{\circ}$	0.18100	**0.14466**$^{\circ}$	**0.09550**$^{\bullet}$	**0.09311**$^{\bullet}$	294	**0.02601**$^{\bullet}$	**0.49559**	0.02100	0.02166	0.03160	**0.02931**$^{\circ}$
	BM25	**258**	0.01851	0.30709	0.11400	0.08766	0.05800	0.06096	**283**	0.01687	0.30631	**0.02900**$^{\bullet}$	0.02900	0.03020	0.02368
Majority @ 20%		1	5	1	-	5	7	4	2	4	-	2	1	2	4
30%	CVV	440	0.02832	0.53794	**0.18400**	0.13366	0.08840	0.08953	428	0.02199	0.45441	0.01800	0.01933	0.02430	0.02355
	Taily	493	0.02909	**0.54828**	0.15700	0.12633	0.09020	0.08729	491	**0.02564**	**0.50108**	0.02200	0.02000	0.02510	0.02772
	CORI	440	**0.02929**	0.53419	0.16800	**0.13700**	**0.09110**	0.08557	**427**	0.02479	0.47308	0.01900	0.02033	0.02680	0.02638
	KL	564	0.01922	0.31722	0.14700	0.09999	0.06590	0.05050	528	0.01772	0.29294	**0.02300**	0.02300	0.03250	0.02033
	vGOISS	**426**	0.02716	0.52556	0.17400	0.13233	0.08750	**0.09071**	**427**	0.02379	0.45443	0.02000	**0.02466**	**0.03520**	**0.02779**
	RW	681	0.02166	0.36004	0.13600	0.10766	0.06980	0.06526	595	0.02003	0.31353	0.01900	0.02433	0.02870	0.01755
	BB2	441	**0.02965**$^{\circ}$	**0.55172**$^{\bullet}$	0.17000	**0.13900**	**0.09270**$^{\bullet}$	0.08895	429	0.02453	0.48468	0.01900	0.02000	0.02660	0.02653
	In_expB2	442	**0.02930**	0.54533	0.16900	**0.13933**	**0.09210**	0.08774	430	0.02425	0.48019	0.01800	0.02000	0.02660	0.02598
	In_expC2	442	**0.02940**	0.54715	0.17100	0.14000	**0.09170**	0.08764	430	0.02434	0.48088	0.01800	0.02000	0.02660	0.02602
	InL2	441	**0.02981**$^{\circ}$	0.52955	0.17400	**0.13700**	**0.09310**$^{\bullet}$	0.08873	427	0.02477	0.48307	0.01900	0.02033	0.02690	0.02676
	Hiemstra_LM	439	**0.02994**$^{\circ}$	0.52976	0.17100	0.14467	**0.09340**$^{\bullet}$	**0.09413**$^{\bullet}$	**425**	**0.02757**$^{\bullet}$	0.49884	0.02100	0.02233	0.03520	**0.03101**
	DFI0	432	**0.02964**	0.53454	**0.19500**$^{\bullet}$	**0.15000**$^{\bullet}$	**0.09550**$^{\bullet}$	**0.10564**$^{\bullet}$	**416**	0.02709	0.49944	0.02000	0.02366	**0.03560**	**0.03134**$^{\bullet}$
	TF_IDF	440	**0.02996**$^{\bullet}$	**0.55905**$^{\bullet}$	0.17600	**0.13833**	**0.09360**$^{\bullet}$	**0.09144**$^{\circ}$	427	0.02484	0.49217	0.01800	0.02000	0.02720	0.02686
	BM25	**392**	0.01871	0.33018	0.12300	0.09267	0.06030	0.05889	**425**	0.01752	0.31956	**0.02300**	0.02367	0.02800	0.02376
Majority @ 30%		1	7	2	1	5	7	3	3	2	-	/	-	1	2
40%	CVV	568	0.02862	0.54157	0.16500	0.13000	0.08930	0.08391	555	0.02153	0.43317	0.01900	0.01833	0.02330	0.02288
	Taily	637	0.02904	**0.55292**	0.14600	0.12267	0.08840	**0.08732**	635	**0.02529**	**0.49908**	**0.02200**	0.02033	0.02420	**0.02673**
	CORI	568	**0.02948**	0.55202	**0.16900**	**0.13366**	**0.09040**	0.08673	**554**	0.02392	0.47566	0.01800	0.01933	0.02500	0.02513
	KL	708	0.02240	0.39749	0.15300	0.11233	0.07340	0.06295	665	0.02006	0.36329	0.01900	0.02133	**0.02950**	0.02259
	vGOISS	**556**	0.02811	0.53928	0.16100	0.12900	0.08680	0.08503	555	0.02428	0.46953	0.01900	0.02100	0.02690	0.02627
	RW	801	0.02367	0.40191	0.13600	0.11867	0.07430	0.05530	726	0.02045	0.32883	0.02000	**0.02300**	0.02630	0.02092
	BB2	570	**0.02972**	**0.56335**$^{\bullet}$	0.16800	**0.13400**	**0.09160**	0.08695	554	0.02415	0.48449	0.01800	0.01933	0.02480	0.02536
	In_expB2	571	**0.02955**	**0.56161**$^{\bullet}$	0.16700	**0.13433**	**0.09150**	0.08619	555	0.02416	0.48381	0.01800	0.01900	0.02450	0.02531
	In_expC2	571	**0.02958**	**0.55682**	0.16800	**0.13467**	**0.09160**	0.08668	554	0.02421	0.48361	0.01800	0.01900	0.02480	0.02535
	InL2	569	**0.02979**$^{\circ}$	0.55044	0.16400	0.13267	**0.09160**	0.08505	**553**	0.02425	0.47810	0.01800	0.01933	0.02500	0.02522
	Hiemstra_LM	568	**0.03007**$^{\bullet}$	0.54696	**0.17800**$^{\bullet}$	**0.14033**$^{\bullet}$	**0.09420**$^{\bullet}$	**0.09224**$^{\bullet}$	**552**	**0.02638**$^{\bullet}$	**0.50025**	0.01800	0.02000	0.02670	**0.02756**$^{\bullet}$
	DFI0	560	**0.02961**	0.53484	**0.17800**$^{\bullet}$	**0.14367**$^{\bullet}$	**0.09420**$^{\bullet}$	**0.10264**$^{\bullet}$	**541**	**0.02648**$^{\bullet}$	0.49494	0.01900	0.02266	**0.03010**	**0.02944**$^{\bullet}$
	TF_IDF	570	**0.02980**$^{\circ}$	**0.56265**$^{\bullet}$	0.16500	**0.13434**	**0.09170**$^{\circ}$	0.08594	554	0.02417	0.48283	0.01800	0.01900	0.02470	0.02523
	BM25	**520**	0.01904	0.33731	0.12100	**0.09266**	0.06240	0.05860	560	0.01801	0.33006	0.02000	0.02267	0.02730	0.02334
Majority @ 40%		1	7	4	2	6	7	2	3	2	1	-	-	1	2
50%	CVV	692	0.02889	0.54909	**0.16500**	**0.13066**	0.08890	0.08371	679	0.02126	0.43806	0.01800	0.01800	0.02320	0.02268
	Taily	776	0.02895	**0.55308**	0.14400	0.12333	0.08780	**0.08648**	773	**0.02503**	**0.49994**	**0.02100**	0.01967	0.02380	**0.02610**
	CORI	693	**0.02930**	0.55217	0.16000	0.12966	**0.08970**	0.08468	**676**	0.02324	0.46961	0.01800	0.01900	0.02380	0.02434
	KL	842	0.02472	0.44224	0.16100	0.11666	0.07840	0.06798	791	0.02152	0.39351	0.01900	0.02066	**0.02770**	0.02336
	vGIOSS	**682**	0.02854	0.54106	0.14500	0.12700	0.08690	0.08155	678	0.02332	0.46587	0.01800	0.01933	0.02420	0.02421
	RW	894	0.02419	0.45144	0.13800	0.11966	0.07770	0.07167	833	0.01921	0.34501	0.01400	**0.02067**	0.02320	0.01660
	BB2	695	**0.02944**	**0.55672**$^{\circ}$	0.15400	**0.13233**	**0.09060**	0.08422	676	0.02346	0.47107	0.01800	0.01900	0.02360	0.02444
	In_expB2	695	**0.02935**	**0.55628**	0.15400	**0.13233**	**0.09010**	0.08408	676	0.02345	0.47101	0.01800	0.01900	0.02360	0.02446
	In_expC2	695	**0.02935**	**0.55485**	0.15100	**0.13266**	**0.09060**	0.08361	676	0.02344	0.47209	0.01800	0.01866	0.02360	0.02447
	InL2	693	**0.02943**	**0.55774**$^{\circ}$	0.15800	**0.13166**	**0.09030**	0.08465	**675**	0.02357	0.47070	0.01800	0.01900	0.02360	0.02436
	Hiemstra_LM	693	**0.02985**$^{\circ}$	0.54513	**0.17100**$^{\bullet}$	**0.13800**$^{\circ}$	**0.09210**$^{\bullet}$	**0.08935**$^{\bullet}$	**673**	0.02465	0.48919	0.01800	0.01933	0.02410	0.02526
	DFI0	691	**0.02953**	0.53339	**0.17000**$^{\bullet}$	**0.13800**$^{\circ}$	**0.09200**$^{\bullet}$	**0.09320**$^{\bullet}$	**667**	**0.02568**	**0.49124**	0.01700	**0.02200**$^{\bullet}$	**0.02770**	**0.02801**$^{\bullet}$
	TF_IDF	695	**0.02937**	**0.55600**	0.15700	**0.13200**	**0.09090**	0.08429	677	0.02337	0.46941	0.01800	0.01900	0.02340	0.02425
	BM25	**646**	0.01986	0.35194	0.12400	0.09433	0.06390	0.06346	692	0.01880	0.35798	**0.02000**	**0.02133**$^{\bullet}$	0.02850	0.02365
Majority @ 50%		1	7	5	2	7	7	2	3	1	-	-	2	/	1

Table A.5: Document Retrieval effectiveness and efficiency on ASIS* Testbeds ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t test).

ASIS*		ASISWOR Testbed							ASISWR Testbed						
%	Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
100%	Flooding	4963	0.02532	0.46296	0.16400	0.11966	0.07620	0.07122	5279	0.01635	0.33847	0.01414	0.01818	0.01899	0.01732
5%	CVV	**420**	0.02384	0.43915	0.16000	0.12200	0.07530	0.06940	**443**	0.02121	0.41605	0.01600	0.02167	0.02390	0.02131
	Taily	548	**0.02577**	**0.44469**	0.15400	0.12267	**0.07820**	0.07047	565	0.02209	0.41179	0.02000	0.02200	0.02590	0.02338
	CORI	433	0.02566	0.43851	**0.16500**	**0.12466**	0.07660	**0.07411**	455	**0.02316**	**0.42360**	0.01700	0.02333	0.02860	0.02351
	KL	593	0.01174	0.18069	0.10900	0.07433	0.04200	0.03286	593	0.01155	0.18331	0.02100	0.03300	0.03590	0.01559
	vGIOSS	488	0.01589	0.29820	0.12300	0.09333	0.05350	0.05675	501	0.01713	0.32008	0.02000	**0.03933**	**0.04390**	**0.02694**
	RW	535	0.01185	0.10320	0.07800	0.05400	0.02920	0.01721	570	0.01309	0.10528	**0.02300**	0.02700	0.02310	0.00855
	BB2	435	**0.02668**$^\bullet$	**0.45975**$^\bullet$	**0.18400**$^\bullet$	**0.14033**$^\bullet$	**0.08170**$^\bullet$	**0.08005**$^\bullet$	459	**0.02337**	**0.43016**	0.01800	0.02467	0.03030	0.02605
	In_expB2	430	0.02500	0.41463	**0.18400**$^\bullet$	**0.13799**$^\bullet$	**0.07990**$^\circ$	0.07214	452	**0.02352**	0.41349	0.01900	0.02633	0.03260	0.02692
	In_expC2	428	0.02471	0.40928	**0.18400**$^\bullet$	**0.13433**$^\bullet$	**0.07960**	0.07108	451	**0.02332**	0.41151	0.02000	0.02700	0.03370	**0.02722**$^\bullet$
	InL2	442	**0.02692**$^\bullet$	**0.45886**$^\bullet$	**0.18000**$^\bullet$	**0.13967**$^\bullet$	**0.08280**$^\bullet$	**0.07812**$^\bullet$	465	**0.02479**$^\bullet$	**0.44998**$^\bullet$	0.01900	0.02500	0.03170	0.02684
	Hiemstra_LM	499	0.02378	0.40370	**0.19000**$^\bullet$	**0.13033**$^\bullet$	**0.08250**$^\bullet$	0.07376	509	**0.02413**$^\circ$	0.41508	0.02000	**0.04200**$^\bullet$	**0.05380**$^\bullet$	**0.03472**$^\bullet$
	DFI0	**364**	0.01234	0.05837	0.08000	0.04633	0.02320	0.01149	**437**	**0.02530**$^\bullet$	**0.43666**	**0.02600**$^\bullet$	0.03333	0.03940	**0.03077**$^\bullet$
	TF_IDF	612	0.01235	0.19098	0.10700	0.07567	0.04250	0.03354	461	**0.02481**$^\bullet$	**0.43865**$^\circ$	0.01900	0.02500	0.03100	0.02630
	BM25	468	0.02393	0.41636	0.15600	0.11900	0.07580	0.07087	487	0.02182	0.40511	0.01600	0.02133	0.02510	0.02297
Majority @ 5%		1	2	2	5	5	5	2	1	7	4	1	1	1	3
20%	CVV	**1499**	0.02554	0.46542	0.16500	0.11966	0.07670	0.07184	**1595**	0.01942	0.38981	0.01400	0.01867	0.02060	0.01864
	Taily	2636	0.02608	0.46309	0.16000	0.11966	0.07830	0.07104	2640	0.01946	0.37962	**0.01900**	0.01966	0.02180	0.02040
	CORI	1553	**0.02632**	**0.47639**	0.16600	**0.12266**	**0.07870**	0.07484	1639	0.01859	0.37742	0.01300	0.01800	0.02060	0.01830
	KL	2065	0.01854	0.35115	0.13400	0.10066	0.06300	0.05238	2101	0.01786	0.35171	0.01400	0.01933	0.02100	0.01823
	vGIOSS	1723	0.02357	0.42905	**0.17300**	0.12233	0.07690	**0.07549**	1792	**0.02205**	**0.42937**	0.01300	0.02033	**0.02370**	**0.02182**
	RW	1803	0.01708	0.24551	0.12900	0.09499	0.05470	0.03257	1918	0.01657	0.24774	0.01800	**0.02233**	0.02050	0.01336
	BB2	1552	**0.02650**	0.47629	0.16900	**0.12833**	**0.08110**$^\bullet$	**0.07613**$^\circ$	1641	0.01873	0.37554	0.01500	0.01833	0.02030	0.01845
	In_expB2	1558	0.02583	0.46643	0.17300	**0.12800**	**0.08020**	0.07517	1649	0.01905	0.37373	0.01500	0.01833	0.02050	0.01863
	In_expC2	1555	0.02572	0.46340	**0.17400**	**0.12766**	**0.08010**	0.07493	1646	0.01906	0.37171	0.01500	0.01833	0.02050	0.01860
	InL2	1572	**0.02650**	**0.47742**	0.16800	**0.12633**	**0.07990**	0.07519	1657	0.01903	0.38115	0.01400	0.01800	0.02030	0.01841
	Hiemstra_LM	1727	**0.02726**$^\bullet$	0.45967	**0.18200**$^\bullet$	**0.13333**$^\bullet$	**0.08360**$^\bullet$	**0.07814**$^\bullet$	1779	**0.02389**$^\bullet$	**0.43013**$^\bullet$	0.01400	0.01800	0.02220	**0.02237**
	DFI0	**1431**	0.01702	0.21208	0.11100	0.08133	0.04530	0.03423	**1616**	**0.02308**$^\bullet$	0.42478	**0.01600**	**0.02133**	**0.02390**	**0.02238**$^\bullet$
	TF_IDF	2083	0.01857	0.35171	0.13400	0.10100	0.06310	0.05239	1642	0.01889	0.37968	0.01400	0.01800	0.02020	0.01831
	BM25	1622	0.02573	0.46237	0.16000	0.12000	0.07730	0.07257	1706	0.02024	0.40079	0.01400	0.01767	0.02000	0.01911
Majority @ 20%		1	3	1	2	5	5	2	-	2	1	-	-	1	2
30%	CVV	**2147**	0.02555	0.46549	0.16400	0.11966	0.07640	0.07179	**2271**	0.01853	0.37236	0.01400	0.01800	0.01930	0.01792
	Taily	4152	0.02574	0.46509	0.15900	0.11866	0.07770	0.07013	4122	0.01850	0.36757	0.01600	0.01867	0.02110	0.01948
	CORI	2219	**0.02605**	**0.47195**	0.16500	0.12133	0.07810	0.07386	2331	0.01711	0.35078	0.01400	0.01733	0.01960	0.01736
	KL	2807	0.02100	0.39297	0.14200	0.10666	0.06730	0.05951	2866	0.01889	0.37122	0.01500	0.01933	0.01920	0.01833
	vGOISS	2420	0.02483	0.45125	**0.16900**	**0.12199**	**0.07840**	**0.07903**	2519	**0.02121**	**0.40903**	0.01400	0.01833	0.01970	**0.01971**
	RW	2447	0.02066	0.32899	0.14600	0.10266	0.06210	0.05248	2607	0.01758	0.29430	**0.02100**	**0.02000**	**0.02260**	0.01631
	BB2	2217	**0.02610**	0.47085	0.16500	**0.12599**	**0.07860**	0.07454	2333	0.01723	0.35243	0.01400	0.01800	0.01940	0.01740
	In_expB2	2221	0.02586	0.46576	0.16600	**0.12566**	**0.07890**	0.07425	2339	0.01724	0.35311	0.01400	0.01800	0.01930	0.01748
	In_expC2	2219	0.02579	0.46490	0.16500	**0.12433**	**0.07890**	0.07392	2338	0.01727	0.35326	0.01400	0.01800	0.01940	0.01749
	InL2	2243	0.02602	**0.47147**	0.16600	**0.12399**	**0.07840**	0.07402	2351	0.01737	0.35631	0.01400	0.01767	0.01950	0.01740
	Hiemstra_LM	2416	**0.02669**	0.46054	**0.17400**	**0.12933**$^\circ$	**0.08170**$^\bullet$	**0.07672**	2498	**0.02155**	0.40213	0.01400	0.01800	0.02010	0.01957
	DFI0	**2065**	0.01823	0.27837	0.12600	0.09533	0.05580	0.04358	**2286**	**0.02126**	0.40267	**0.01700**	**0.02100**	**0.02180**	**0.02048**$^\circ$
	TF_IDF	2823	0.02096	0.39228	0.14200	0.10666	0.06730	0.05948	2334	0.01725	0.35450	0.01400	0.01733	0.01930	0.01736
	BM25	2326	0.02559	0.45953	0.16200	0.11966	0.07690	0.07225	2445	0.01940	0.38511	0.01400	0.01733	0.01970	0.01849
Majority @ 30%		1	2	-	1	5	4	-	-	2	-	-	1	-	1
40%	CVV	**2741**	0.02558	0.46588	0.16300	0.11966	0.07640	0.07149	**2892**	0.01741	0.35826	0.01400	0.01800	0.01950	0.01768
	Taily	5609	0.02556	0.46440	0.15800	0.11900	0.07760	0.06978	5547	0.01788	0.36117	0.01500	**0.01900**	0.02070	**0.01876**
	CORI	2815	**0.02587**	**0.46870**	0.16300	0.12033	0.07760	0.07274	2957	0.01638	0.33771	0.01400	0.01733	0.01910	0.01699
	KL	3414	0.02274	0.41503	0.14700	0.11033	0.07190	0.06422	3518	0.01892	0.38250	0.01400	0.01867	0.01900	0.01810
	vGOISS	3029	0.02557	0.45935	**0.17000**	**0.12266**	**0.07930**	**0.07447**	3156	**0.01996**	**0.39114**	0.01400	0.01767	0.01950	0.01871
	RW	2989	0.01976	0.33711	0.13100	0.09900	0.06320	0.04688	3177	0.01834	0.32480	**0.02100**	0.01867	**0.02270**	0.01631
	BB2	2815	**0.02588**	0.46744	0.16400	**0.12266**	0.07800	0.07289	2955	0.01644	0.33825	0.01400	0.01733	0.01900	0.01700
	In_expB2	2821	0.02586	0.46753	0.16400	**0.12300**	0.07820	0.07308	2962	0.01645	0.33801	0.01400	0.01767	0.01910	0.01699
	In_expC2	2820	0.02576	0.46464	0.16400	0.12300	0.07850	0.07311	2962	0.01646	0.33831	0.01400	0.01767	0.01910	0.01700
	InL2	2842	**0.02588**	0.46851	0.16400	0.12100	0.07790	0.07260	2978	0.01640	0.33791	0.01400	0.01733	0.01900	0.01697
	Hiemstra_LM	3019	**0.02656**$^\bullet$	0.46210	**0.16900**	**0.12533**	**0.08080**	**0.07570**$^\circ$	3132	0.01837	0.37108	0.01400	0.01733	0.01910	0.01757
	DFI0	**2666**	0.02067	0.33768	0.14100	0.10233	0.06400	0.05713	**2872**	**0.01962**	**0.38610**	**0.01600**	**0.02033**	**0.02090**	**0.01917**$^\bullet$
	TF_IDF	3426	0.02273	0.41475	0.14700	0.11033	0.07190	0.06422	2959	0.01630	0.33689	0.01400	0.01733	0.01900	0.01694
	BM25	2981	0.02562	0.46525	0.16200	0.11900	0.07660	0.07198	3112	0.01864	0.37773	0.01400	0.01733	0.01980	0.01811
Majority @ 40%		1	3	-	-	3	1	1	1	-	-	-	1	-	1
50%	CVV	**3289**	0.02549	0.46473	0.16300	0.11966	0.07640	0.07144	**3465**	0.01749	0.35950	0.01400	0.01767	0.01900	0.01743
	Taily	7041	0.02556	**0.46792**	0.16000	0.11900	0.07760	0.07035	6978	0.01726	0.35735	0.01400	**0.01933**	**0.02040**	**0.01832**
	CORI	3358	0.02576	0.46707	0.16400	0.12033	0.07730	0.07222	3531	0.01589	0.32790	0.01400	0.01733	0.01910	0.01680
	KL	3893	0.02353	0.43121	0.15500	0.11399	0.07400	0.06642	4047	0.01890	0.37304	0.01400	0.01800	0.01870	0.01779
	vGIOSS	3562	**0.02588**	0.46205	**0.16800**	**0.12333**	**0.07980**	**0.07403**	3724	**0.01907**	**0.38323**	0.01400	0.01767	0.01890	0.01807
	RW	3442	0.02211	0.37198	0.13900	0.10600	0.06680	0.05370	3657	0.01703	0.29759	**0.02000**	0.01800	0.02010	0.01548
	BB2	3357	0.02576	0.46642	0.16400	0.12133	0.07770	0.07244	3529	0.01591	0.32844	0.01400	0.01733	0.01900	0.01681
	In_expB2	3361	**0.02577**	0.46630	0.16400	0.12133	0.07780	0.07247	3535	0.01588	0.32769	0.01400	0.01767	0.01900	0.01678
	In_expC2	3358	**0.02577**	0.46679	**0.16500**	**0.12199**	**0.07790**	**0.07285**	3533	0.01589	0.32784	0.01400	0.01767	0.01900	0.01678
	InL2	3801	0.02366	0.43688	0.15800	0.11566	0.07450	0.06768	3548	0.01591	0.32790	0.01400	0.01733	0.01900	0.01680
	Hiemstra_LM	3901	0.02353	0.43121	0.15500	0.11399	0.07400	0.06642	3698	0.01702	0.34692	0.01400	0.01733	0.01900	0.01702
	DFI0	**3266**	0.02215	0.38501	0.14500	0.10833	0.06780	0.06169	**3395**	0.01853	0.36957	0.01500	**0.01933**	**0.02070**	**0.01829**
	TF_IDF	3901	0.02353	0.43121	0.15500	0.11399	0.07400	0.06642	3530	0.01586	0.32715	0.01400	0.01733	0.01910	0.01678
	BM25	3550	0.02556	0.46553	0.16100	0.11800	0.07630	0.07173	3699	0.01813	0.36748	0.01400	0.01733	0.01960	0.01789
Majority @ 50%		1	-	-	-	-	-	-	1	-	-	-	/	1	-

Table A.6: Document Retrieval effectiveness and efficiency on U* Testbeds ($\circ$ & $\bullet$ indicate statistical significance at $p < 0.05$ and $p < 0.01$ respectively using bootstrapping two-paired t test).

U*		UWOR Testbed							UWR Testbed						
%	Method	#Peers	Precision	Recall	P@10	P@30	P@100	MAP	#Peers	Precision	Recall	P@10	P@30	P@100	MAP
100%	Flooding	2499	0.02764	0.49910	0.21200	0.14800	0.09580	0.10581	1337	0.02269	0.42657	0.01400	0.01900	0.02450	0.02331
5%	CVV	657	0.01767	0.30214	0.17500	0.11333	0.06420	0.06646	732	0.01697	0.30835	0.01200	0.01933	0.02270	0.01708
	Taily	847	**0.02589**	**0.43444**	**0.18000**	**0.13966**	**0.08770**	**0.08986**	860	**0.02470**	**0.42483**	0.01700	0.02500	0.03150	0.02852
	CORI	608	0.02395	0.40440	0.17600	0.13033	0.08540	0.07654	661	0.02434	0.42469	0.01700	0.03433	**0.04720**	**0.03239**
	KL	841	0.01047	0.13661	0.10200	0.06633	0.03330	0.03441	1037	0.01005	0.16173	0.01900	0.02333	0.02370	0.01194
	vGIOSS	**605**	0.01490	0.24839	0.13500	0.09066	0.05190	0.04225	**648**	0.01556	0.28677	0.01800	**0.03667**	0.04140	0.02018
	RW	1117	0.01565	0.08328	0.11500	0.06633	0.03180	0.01959	1327	0.01392	0.11845	**0.02900**	0.02666	0.01920	0.00804
	BB2	627	**0.02887**$^\bullet$	**0.48547**$^\bullet$	**0.22600**$^\bullet$	**0.16367**$^\bullet$	**0.10150**$^\bullet$	**0.10895**$^\bullet$	687	**0.02834**$^\bullet$	**0.49723**$^\bullet$	0.02600	**0.03767**$^\bullet$	**0.05270**$^\bullet$	**0.03943**$^\bullet$
	In_expB2	622	0.02248	0.40506	**0.20700**$^\bullet$	**0.13967**	0.08080	0.08803	684	0.02276	0.40535	**0.02900**	**0.04433**$^\bullet$	**0.05220**$^\bullet$	**0.03247**
	In_expC2	620	0.02185	0.39718	**0.19700**$^\bullet$	0.13733	0.07930	0.08608	680	0.02208	0.39184	0.02800	**0.04333**$^\bullet$	**0.05110**$^\bullet$	0.03111
	InL2	**604**	**0.03016**$^\bullet$	**0.50283**$^\bullet$	**0.20800**$^\bullet$	**0.16166**$^\bullet$	**0.10410**$^\bullet$	**0.10537**$^\bullet$	658	**0.03018**$^\bullet$	**0.51054**$^\bullet$	0.02500	**0.04099**$^\bullet$	**0.05970**$^\bullet$	**0.04377**$^\bullet$
	Hiemstra_LM	**595**	**0.02889**$^\bullet$	**0.48342**$^\bullet$	**0.21800**$^\bullet$	**0.16100**$^\bullet$	**0.10190**$^\bullet$	**0.10393**$^\bullet$	**643**	**0.02926**$^\bullet$	**0.50227**$^\bullet$	0.03100	**0.05466**$^\bullet$	**0.06570**$^\bullet$	**0.04592**$^\bullet$
	DFI0	**603**	**0.02745**$^\bullet$	**0.47140**$^\bullet$	**0.22600**$^\bullet$	**0.16299**$^\bullet$	**0.10060**$^\bullet$	**0.10666**$^\bullet$	660	**0.02762**$^\bullet$	**0.48452**$^\bullet$	0.02900	**0.04533**$^\bullet$	**0.05840**$^\bullet$	**0.04035**$^\bullet$
	TF_IDF	608	**0.03051**$^\bullet$	**0.51930**$^\bullet$	**0.21800**$^\bullet$	**0.16800**$^\bullet$	**0.10570**$^\bullet$	**0.11207**$^\bullet$	660	**0.03044**$^\bullet$	**0.52585**$^\bullet$	0.02400	**0.04133**$^\bullet$	**0.06150**$^\bullet$	**0.04430**$^\bullet$
	BM25	**598**	0.01996	0.34450	0.15600	0.11733	0.07210	0.08672	**588**	0.01971	0.33656	**0.03800**$^\bullet$	**0.03967**$^\bullet$	0.04380	**0.03471**$^\circ$
Majority @ 5%		4	5	5	7	6	5	5	2	5	5	2	7	7	6
20%	CVV	2363	0.02520	0.45742	0.19300	**0.14667**	0.09100	0.09745	2563	0.02080	0.39984	0.01300	0.01767	0.02160	0.02061
	Taily	3447	0.02862	**0.50309**	0.19700	0.14499	0.09710	**0.10476**	3487	0.02467	0.45741	0.01400	0.01767	0.02580	**0.02504**
	CORI	2204	**0.02914**	0.50222	**0.19800**	0.14500	**0.09850**	0.10345	2364	**0.02540**	**0.47351**	0.01200	0.01600	0.02300	0.02500
	KL	3027	0.01727	0.29480	0.15400	0.10866	0.06350	0.06229	3608	0.01731	0.32185	0.01300	0.01900	0.02320	0.01766
	vGIOSS	**2193**	0.02255	0.42168	0.17800	0.12867	0.08170	0.08684	**2326**	0.02203	0.42049	0.01200	**0.02333**	**0.03160**	0.02493
	RW	3581	0.01862	0.24733	0.15500	0.10900	0.06080	0.04517	4090	0.01743	0.25501	**0.01900**	0.02167	0.02160	0.01265
	BB2	2244	**0.03001**	**0.52063**$^\bullet$	**0.21100**$^\bullet$	**0.15033**$^\bullet$	**0.10080**$^\circ$	**0.10639**$^\circ$	2422	**0.02604**$^\bullet$	**0.48143**$^\circ$	0.01500	0.01800	0.02420	**0.02591**$^\circ$
	In_expB2	2249	0.02653	0.46158	**0.21900**$^\bullet$	**0.14900**$^\circ$	0.09320	0.09736	2427	0.02375	0.44453	0.01500	0.01833	0.02450	0.02358
	In_expC2	2248	0.02605	0.45686	**0.21600**$^\bullet$	**0.14800**$^\circ$	0.09150	0.09648	2425	0.02367	0.44414	0.01500	0.01867	0.02440	0.02351
	InL2	2196	**0.03099**	**0.53805**$^\bullet$	**0.20600**	0.14366	**0.10120**$^\circ$	**0.10620**$^\circ$	2356	**0.02730**$^\bullet$	**0.49972**$^\bullet$	0.01400	0.01667	0.02460	**0.02638**$^\bullet$
	Hiemstra_LM	**2158**	**0.03048**	**0.52333**$^\bullet$	**0.20100**	**0.14933**	**0.10020**$^\circ$	**0.10490**	**2312**	**0.02821**$^\bullet$	**0.51067**$^\bullet$	0.01500	0.01966	0.02830	**0.02933**$^\bullet$
	DFI0	**2177**	**0.02947**	**0.50665**	**0.20700**$^\bullet$	**0.15367**$^\bullet$	**0.10020**$^\circ$	0.10406	2343	**0.02684**$^\bullet$	**0.49642**$^\circ$	0.01400	0.01800	0.02580	**0.02724**$^\bullet$
	TF_IDF	2207	**0.03108**$^\bullet$	**0.54088**$^\bullet$	**0.20700**$^\bullet$	0.14500	**0.10120**$^\circ$	**0.10712**$^\circ$	2368	**0.02728**$^\bullet$	**0.49583**$^\circ$	0.01400	0.01667	0.02450	**0.02611**$^\bullet$
	BM25	**2159**	0.02295	0.42481	0.16500	0.12233	0.08160	0.10031	**2152**	0.02163	0.40981	0.01400	0.01900	0.02750	**0.02586**
Majority @ 20%		3	5	5	7	5	5	4	2	5	5	-	-	-	6
30%	CVV	3371	0.02704	0.47601	0.19300	**0.14333**	0.09300	0.10035	3641	0.02143	0.41303	0.01300	0.01700	0.02060	0.02096
	Taily	5139	0.02860	0.51461	0.18800	0.14000	0.09570	0.10377	5190	0.02347	0.44768	0.01300	0.01633	0.02380	0.02323
	CORI	3180	**0.02960**	**0.51962**	**0.20200**	0.14300	**0.09880**	**0.10423**	3404	**0.02445**	**0.47142**	0.01300	0.01567	0.02150	0.02367
	KL	4173	0.02014	0.34941	0.16500	0.11866	0.07420	0.07051	4771	0.01963	0.37313	0.01300	0.01767	0.02210	0.01962
	vGOISS	**3168**	0.02506	0.45716	0.18300	0.13867	0.08860	0.09153	**3361**	0.02332	0.43796	**0.01400**	**0.01900**	**0.02420**	**0.02414**
	RW	4697	0.02177	0.31995	0.17300	0.11767	0.07030	0.05665	5260	0.01819	0.30229	0.01200	0.01767	0.01890	0.01379
	BB2	3224	**0.03006**	**0.52275**$^\bullet$	**0.20800**	**0.14833**	**0.09940**	**0.10565**	3464	0.02424	0.46139	0.01300	0.01667	0.02180	0.02357
	In_expB2	3236	0.02742	0.48868	**0.21900**$^\bullet$	**0.14933**$^\circ$	0.09430	0.09938	3478	0.02279	0.43930	0.01300	0.01700	0.02210	0.02216
	In_expC2	3234	0.02711	0.48095	**0.22000**$^\bullet$	**0.15100**$^\bullet$	0.09430	0.09906	3473	0.02265	0.43502	0.01300	0.01700	0.02220	0.02209
	InL2	3170	**0.03040**	**0.53276**$^\bullet$	0.20100	0.14100	0.09640	0.10374	3391	**0.02477**	**0.47328**	0.01200	0.01567	0.02150	0.02351
	Hiemstra_LM	**3116**	**0.03063**	**0.53165**$^\bullet$	**0.20400**	**0.14566**	0.09820	**0.10482**	**3338**	**0.02647**$^\bullet$	**0.48765**$^\bullet$	**0.01400**	0.01667	0.02190	**0.02520**$^\bullet$
	DFI0	5048	0.01500	0.25000	0.20000	0.13330	0.07000	0.05000	**3351**	**0.02572**$^\bullet$	**0.47891**	**0.01400**	0.01667	0.02240	**0.02462**
	TF_IDF	3186	**0.03042**	**0.53484**$^\bullet$	0.20100	0.14167	0.09720	0.10365	3412	**0.02477**	**0.47417**	0.01200	0.01567	0.02120	0.02338
	BM25	**3089**	0.02356	0.43074	0.17100	0.12233	0.08080	0.09985	**3088**	0.02110	0.40423	0.01300	**0.01800**	**0.02490**	0.02391
Majority @ 30%		2	4	4	4	4	1	2	3	4	4	/	-	1	2
40%	CVV	4291	0.02806	0.49161	**0.20000**	**0.14033**	0.09380	0.09972	4607	0.02123	0.41916	0.01300	0.01700	0.02010	0.02065
	Taily	6650	0.02855	**0.51672**	0.18800	0.13700	0.09290	**0.10322**	6676	0.02229	0.42539	0.01400	0.01600	0.02180	0.02184
	CORI	4109	**0.02952**	0.51493	0.19200	0.14000	**0.09560**	0.10277	4381	**0.02329**	**0.45334**	0.01200	0.01567	0.02070	0.02226
	KL	5226	0.02256	0.39250	0.18700	0.12633	0.07980	0.07919	5805	0.02022	0.38918	0.01300	0.01700	0.02070	0.02000
	vGOISS	**4094**	0.02640	0.47397	0.19000	0.13833	0.09110	0.09232	**4334**	0.02300	0.43966	0.01400	0.01767	**0.02190**	**0.02290**
	RW	5545	0.02227	0.37390	0.17500	0.11967	0.07550	0.06813	6103	0.01956	0.31970	**0.01500**	**0.02033**	0.02030	0.01650
	BB2	4150	**0.02962**	0.51072	**0.20500**	**0.14400**	**0.09730**$^\bullet$	**0.10355**	4437	0.02299	0.44600	**0.01300**	**0.01667**	0.02090	0.02221
	In_expB2	4158	0.02791	0.48941	**0.20700**	**0.14367**	0.09290	0.09827	4450	0.02224	0.43416	**0.01300**	**0.01667**	0.02060	0.02124
	In_expC2	4155	0.02780	0.48823	**0.21000**$^\bullet$	**0.14500**$^\circ$	0.09300	0.09836	4444	0.02212	0.43346	**0.01300**	**0.01667**	0.02080	0.02121
	InL2	4096	**0.02980**	**0.51972**$^\circ$	0.19400	0.13800	0.09420	0.10220	4366	0.02358	**0.46169**$^\bullet$	0.01200	0.01567	0.02070	0.02198
	Hiemstra_LM	**4037**	**0.03015**$^\bullet$	**0.51926**$^\circ$	**0.20400**	**0.14300**	**0.09670**$^\bullet$	**0.10341**	**4306**	**0.02444**$^\bullet$	**0.46686**$^\bullet$	0.01200	0.01567	0.02120	0.02287
	DFI0	**3988**	**0.03005**$^\bullet$	**0.52123**$^\bullet$	**0.20600**	**0.14733**$^\circ$	**0.09800**$^\bullet$	**0.10427**	**4287**	**0.02429**$^\bullet$	**0.46216**$^\bullet$	**0.01300**	**0.01667**	0.02190	**0.02296**
	TF_IDF	4122	**0.02982**	**0.51944**$^\circ$	0.19600	0.13900	0.09440	0.10231	4396	**0.02342**	**0.46087**$^\bullet$	0.01200	0.01567	0.02070	0.02188
	BM25	**4004**	0.02376	0.43412	0.17300	0.12400	0.08190	0.09573	**3993**	0.02077	0.40547	**0.01300**	0.01600	**0.02220**$^\bullet$	0.02280
Majority 40%		3	5	4	5	5	3	3	3	3	4	-	-	1	1
50%	CVV	5144	0.02854	0.49738	**0.20300**	**0.14033**	**0.09440**	0.09896	5496	0.02095	0.41891	0.01200	0.01633	0.02000	0.02039
	Taily	8062	0.02833	**0.51341**	0.19100	0.13800	0.09220	**0.10280**	8047	0.02165	0.42222	0.01200	0.01600	**0.02060**	0.02095
	CORI	4981	**0.02940**	0.50998	0.19700	0.13833	**0.09440**	0.10172	5305	0.02213	0.43374	0.01200	0.01567	0.02010	0.02129
	KL	6144	0.02490	0.42939	0.18300	0.13300	0.08550	0.08557	6690	0.02126	0.41381	0.01200	0.01600	0.02020	0.02055
	vGIOSS	**4969**	0.02743	0.48502	0.19700	0.13533	0.09000	0.09404	**5258**	**0.02222**	**0.43483**	0.01300	0.01567	**0.02060**	**0.02135**
	RW	6191	0.02438	0.38815	0.18600	0.12667	0.08210	0.06903	6732	0.01956	0.34757	**0.01400**	**0.01833**	0.01990	0.01710
	BB2	5015	0.02924	0.50341	0.20100	**0.14033**	**0.09500**$^\bullet$	0.10202	5345	0.02192	0.42836	**0.01300**	0.01600	0.02030	0.02133
	In_expB2	5022	0.02817	0.48882	0.20000	**0.14100**$^\circ$	0.09370	0.09846	5352	0.02153	0.42251	**0.01300**	0.01600	0.02040	0.02083
	In_expC2	5021	0.02816	0.48941	0.20100	**0.14166**	0.09340	0.09875	5351	0.02147	0.42149	**0.01300**	0.01600	0.02040	0.02076
	InL2	**4967**	0.02930	0.51150	0.19100	0.13733	0.09290	0.10111	5283	**0.02251**	**0.43957**$^\circ$	0.01200	0.01567	0.02020	0.02110
	Hiemstra_LM	**4914**	**0.02962**	0.51206	0.19800	**0.14033**	**0.09590**$^\bullet$	0.10236	**5230**	**0.02303**$^\bullet$	**0.44521**$^\bullet$	0.01200	0.01567	0.02040	**0.02150**
	DFI0	**4832**	**0.02963**	**0.51467**	**0.20200**	**0.14366**$^\bullet$	**0.09600**$^\bullet$	**0.10288**	**5219**	**0.02328**$^\bullet$	**0.44834**$^\bullet$	**0.01300**	0.01600	**0.02130**$^\bullet$	**0.02196**$^\circ$
	TF_IDF	4993	0.02932	0.51201	0.19300	0.13733	0.09360	0.10139	5316	**0.02235**	**0.43793**	0.01200	0.01567	0.02030	0.02105
	BM25	**4884**	0.02411	0.44987	0.17300	0.12467	0.08290	0.09555	**4868**	0.02037	0.40408	0.01200	**0.01633**	**0.02150**$^\bullet$	**0.02192**$^\circ$
Majority @ 50%		4	2	1	-	5	3	1	3	4	4	-	-	2	3

A.3 Learning to Route Approach

Table A.7: LTRo Retrieval effectiveness

DL*	**DLWOR test-bed**						**DLWR test-bed**					
Method	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
Taily	0.02815	0.52157	0.16050	0.12850	0.08825	0.08944	0.02519	0.48203	0.02367	0.02216	0.02840	0.02786
CORI	0.02797	0.49827	0.16967	0.13433	0.08813	0.08450	0.02431	0.45318	0.02417	0.02844	0.03362	0.02822
CVV	0.02657	0.49614	0.17567	0.13077	0.08553	0.08836	0.02162	0.42538	0.01983	0.02038	0.02730	0.02655
BB2	0.02899	0.52380	0.17650	0.14139	0.09273	0.09252	0.02504	0.47208	0.02417	0.02711	0.03430	0.02998
Hiemstra_LM	0.02808	0.48895	0.18050	0.13978	0.09120	0.08963	0.02592	0.45913	0.03567	0.04111	0.04527	0.03486
DFI0	0.02840	0.49983	0.18800	0.14405	0.09280	0.09767	0.02637	0.47194	0.02767	0.03561	0.04322	0.03411
TF-IDF	0.02927	0.52444	0.17917	0.14150	0.09390	0.09112	0.02562	0.47399	0.02533	0.02983	0.03663	0.03066
LR	0.04017	**0.57933**	0.22633	0.17783	**0.11153**	0.13322	**0.03687**	**0.54805**	0.05650	0.05561	0.05142	0.06389
MLP	**0.04028**	0.57889	0.22683	**0.17833**	0.11148	0.13135	0.03679	0.54706	0.05650	**0.05572**	0.05160	0.06389
MART	0.03972	0.56748	0.23000	0.17722	0.11043	0.13457	0.03668	0.53493	**0.05767**	0.05539	0.05258	0.06506
RankNet	0.03972	0.56771	0.23000	0.17739	0.11048	0.13472	0.03667	0.53470	**0.05767**	0.05539	0.05252	0.06503
RankBoost	0.03972	0.56756	0.23033	0.17733	0.11045	0.13482	0.03668	0.53500	**0.05767**	0.05539	0.05258	0.06507
AdaRank	0.03972	0.56756	**0.23050**	0.17733	0.11045	**0.13488**	0.03667	0.53470	**0.05767**	0.05539	0.05252	0.06504
Coordinate Ascent	0.03972	0.56771	0.23000	0.17739	0.11048	0.13472	0.03668	0.53508	**0.05767**	0.05539	0.05262	0.06509
LambdaRank	0.03972	0.56756	0.23033	0.17733	0.11045	0.13483	0.03668	0.53516	**0.05767**	0.05539	**0.05263**	**0.06510**
LambdaMART	0.03972	0.56741	0.23000	0.17717	0.11042	0.13450	0.03668	0.53493	**0.05767**	0.05539	0.05258	0.06506
RandomForests	0.03972	0.56748	0.23000	0.17728	0.11043	0.13458	0.03668	0.53493	**0.05767**	0.05539	0.05260	0.06506
Av. Impro.LR	(+**42.55%**)	(+**14.22%**)	(+**29.09%**)	(+**29.84%**)	(+**23.55%**)	(+**47.51%**)	(+**48.83%**)	(+**18.67%**)	(+**125.26%**)	(+**100.31%**)	(+**49.31%**)	(+**112.62%**)
Av. Impro.MLP	(+**42.91%**)	(+**14.14%**)	(+**29.37%**)	(+**30.21%**)	(+**23.50%**)	(+**45.45%**)	(+**48.52%**)	(+**18.45%**)	(+**125.26%**)	(+**100.72%**)	(+**49.84%**)	(+**112.62%**)
Av. Impro.LTRo	(+**40.95%**)	(+**11.90%**)	(+**31.26%**)	(+**29.46%**)	(+**22.35%**)	(+**49.16%**)	(+**48.06%**)	(+**15.83%**)	(+**129.92%**)	(+**99.52%**)	(+**52.68%**)	(+**116.54%**)
ASIS*	**ASISWOR test-bed**						**ASISWR test-bed**					
Taily	0.02581	0.46064	0.15833	0.12022	0.07798	0.07046	0.01934	0.37833	0.01733	0.02000	0.02220	0.02042
CORI	0.02600	0.46448	0.16400	0.12188	0.07780	0.07368	0.01863	0.36941	0.01433	0.01872	0.02163	0.01885
CVV	0.02518	0.45945	0.16350	0.12050	0.07647	0.07314	0.01910	0.38462	0.01467	0.01889	0.02063	0.01954
BB2	0.02630	0.46885	0.17033	0.12844	0.07972	0.07574	0.01878	0.37130	0.01500	0.01928	0.02188	0.01942
Hiemstra_LM	0.02560	0.44269	0.17267	0.12660	0.08065	0.07418	0.02169	0.40091	0.01567	0.02294	0.02837	0.02323
DFI0	0.01752	0.23113	0.11667	0.08222	0.04863	0.03735	0.02207	0.40961	0.01817	0.02317	0.02600	0.02287
TF-IDF	0.01894	0.33906	0.13417	0.09894	0.06207	0.05312	0.01911	0.37406	0.01500	0.01906	0.02197	0.01943
LR	0.04356	0.54821	0.24500	0.17606	0.10615	0.12134	**0.03965**	0.52124	**0.06400**	0.05900	0.04972	0.06027
MLP	0.04354	0.54891	0.24400	0.17561	0.10598	0.12099	0.03954	0.52066	**0.06400**	0.05900	0.04965	0.06021
MART	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
RankNet	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
RankBoost	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52463**	0.06317	**0.06016**	**0.05060**	**0.06194**
AdaRank	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
Coordinate Ascent	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
LambdaRank	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
LambdaMART	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
RandomForests	**0.04412**	**0.55917**	**0.24600**	**0.17733**	**0.10642**	**0.12454**	0.03959	**0.52462**	0.06317	**0.06016**	**0.05060**	**0.06194**
Av. Impro.LR	(+**89.08%**)	(+**42.28%**)	(+**61.76%**)	(+**57.89%**)	(+**52.21%**)	(+**97.13%**)	(+**100.93%**)	(+**35.9%**)	(+**309.46%**)	(+**192.85%**)	(+**116.4%**)	(+**195.3%**)
Av. Impro.MLP	(+**89%**)	(+**42.47%**)	(+**61.1%**)	(+**57.5%**)	(+**51.97%**)	(+**96.55%**)	(+**100.4%**)	(+**35.76%**)	(+**309.46%**)	(+**192.85%**)	(+**116.1%**)	(+**195%**)
Av. Impro.LTRo	(**91.54%**)	(+**45.13%**)	(+**62.42%**)	(+**52.6%**)	(+**52.6%**)	(+**102.32%**)	(+**100.64%**)	(+**36.8%**)	(+**304.13%**)	(+**198.63%**)	(+**120.23%**)	(+**203.45%**)
U*	**UWOR test-bed**						**UWR test-bed**					
Taily	0.02797	0.49474	0.18783	0.14100	0.09313	0.10229	0.02374	0.43882	0.01400	0.01861	0.02545	0.02451
CORI	0.02815	0.48653	0.19267	0.14022	0.09425	0.09706	0.02439	0.45350	0.01333	0.01966	0.02760	0.02572
CVV	0.02468	0.43409	0.19167	0.13644	0.08582	0.09860	0.02011	0.38681	0.01250	0.01767	0.02123	0.02355
BB2	0.02959	0.50708	0.21083	0.15089	0.09940	0.10555	0.02528	0.47091	0.01617	0.02128	0.02867	0.02745
Hiemstra_LM	0.02997	0.51251	0.20617	0.14911	0.09903	0.10436	0.02693	0.48887	0.01750	0.02511	0.03350	0.03031
DFI0	0.02920	0.50360	0.21067	0.15205	0.09900	0.10462	0.02591	0.47932	0.01733	0.02344	0.03168	0.02852
TF-IDF	0.03036	0.52696	0.20417	0.14733	0.09930	0.10590	0.02636	0.48873	0.01517	0.02116	0.03060	0.02841
LR	0.08842	0.73835	0.47133	0.35867	0.21113	0.32898	0.08488	0.71121	0.12317	0.12383	0.09630	0.14347
MLP	0.08842	0.73908	0.47233	0.35861	0.21110	0.32908	0.08485	0.71135	0.12317	0.12394	0.09628	0.14342
MART	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
RankNet	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
RankBoost	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
AdaRank	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
Coordinate Ascent	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
LambdaRank	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
LambdaMART	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
RandomForests	**0.08856**	**0.74407**	**0.47400**	**0.36056**	**0.21228**	**0.33283**	**0.08498**	**0.71152**	**0.12783**	**0.12572**	**0.09957**	**0.14663**
Av. Impro.LR	(+**210.91%**)	(+**49.64%**)	(+**135.46%**)	(+**147.23%**)	(+**121.18%**)	(+**220.9%**)	(+**246.87%**)	(+**56.17%**)	(+**725.31%**)	(+**497.8%**)	(+**246.13%**)	(+**436.64%**)
Av. Impro.MLP	(+**210.92%**)	(+**49.8%**)	(+**135.96%**)	(+**147.2%**)	(+**121.14%**)	(+**221%**)	(+**246.74%**)	(+**56.20%**)	(+**725.31%**)	(+**498.34%**)	(+**246.07%**)	(+**436.46%**)
Av. Impro.LTRo	(+**211.42%**)	(+**50.80%**)	(+**136.80%**)	(+**148.53%**)	(+**122.38%**)	(+**224.65%**)	(+**247.29%**)	(+**56.24%**)	(+**756.58%**)	(+**506.92%**)	(+**257.87%**)	(+**448.5%**)

Appendix B

Reputation-based Query Routing Approaches

B.1 Reputation-based Approaches

Table B.1: Retrieval effectiveness on Digital Library environments (Scenario 1)

DL*		DLWOR test-bed						DLWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	CORI	0.02888	0.53648	0.169	0.13766	0.0915	0.09311	0.01733	0.22004	0.031	0.027	0.0276	0.01587
	Taily	0.02849	0.53061	0.163	0.13	0.0896	0.09006	0.0258	0.5038	0.023	0.02133	0.0271	0.02872
	R	0.02777	0.48951$^{\Leftarrow\Downarrow}$	0.143	**0.13366**	0.0885	0.07800$^{\leftarrow\Downarrow}$	**0.02421**$^{\Rightarrow\downarrow}$	**0.44826**$^{\Rightarrow\Downarrow}$	**0.02800**	**0.02466**	0.02680	**0.02476**$^{\Rightarrow\Downarrow}$
	RT	0.02777	0.48951$^{\Leftarrow\Downarrow}$	0.143	0.13366	0.0885	0.07800$^{\leftarrow\Downarrow}$	**0.02422**$^{\Rightarrow\downarrow}$	**0.44872**$^{\Rightarrow\Downarrow}$	**0.02800**	**0.02466**	0.02690	**0.02482**$^{\Rightarrow\Downarrow}$
	RP	**0.03964**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.66027**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23900**$^{\rightarrow\uparrow\blacktriangle}$	**0.24500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17970**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19164**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03019**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.53787**$^{\Rightarrow\blacktriangle}$	**0.13500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12333**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09370**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08394**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.03964**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.66027**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23900**$^{\rightarrow\uparrow\blacktriangle}$	**0.24500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17970**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19164**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03020**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.53832**$^{\Rightarrow\blacktriangle}$	**0.13500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12333**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09380**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08411**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	CORI	0.02937	0.55104	0.151	0.12733	0.0901	0.08714	0.01978	0.30967	0.025	0.02533	0.0244	0.01974
	Taily	0.02909	0.54828	0.157	0.12633	0.0902	0.08729	0.02564	0.50108	0.022	0.02	0.0251	0.02772
	R	**0.02931**	0.53076	0.147	**0.13566**$^{\Uparrow}$	**0.09170**	0.0835	**0.02495**$^{\Rightarrow}$	**0.47369**$^{\Rightarrow\downarrow}$	**0.02800**$^{\uparrow}$	**0.02233**	**0.02660**	**0.02590**$^{\Rightarrow}$
	RT	0.02932	0.53122	0.146	**0.13600**$^{\Uparrow}$	**0.09180**	0.08389	**0.02496**$^{\Rightarrow}$	**0.47415**$^{\Rightarrow\downarrow}$	**0.02800**$^{\uparrow}$	**0.02233**	**0.02670**	**0.02595**$^{\Rightarrow}$
	RP	**0.04296**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.73606**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19040**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.21174**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03240**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.59779**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11833**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09360**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08804**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04297**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.73652**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25033**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19050**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.21219**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03241**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.59825**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11833**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09370**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08820**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	CORI	0.02972	0.56788	0.157	0.12633	0.0896	0.08718	0.02125	0.3681	0.026	0.026	0.0243	0.02171
	Taily	0.02904	0.55292	0.146	0.12267	0.0884	0.08732	0.02529	0.49908	0.022	0.02033	0.0242	0.02673
	R	**0.02977**$^{\uparrow}$	0.54925$^{\leftarrow}$	0.144	**0.13633**$^{\Rightarrow\Uparrow}$	**0.09160**$^{\uparrow}$	0.0849	**0.02528**$^{\Rightarrow}$	**0.49001**$^{\Rightarrow}$	**0.02500**	**0.02166**	**0.02590**$^{\uparrow}$	**0.02679**$^{\rightarrow}$
	RT	**0.02977**$^{\uparrow}$	0.54925$^{\leftarrow}$	0.144	**0.13633**$^{\Rightarrow\Uparrow}$	**0.09160**$^{\uparrow}$	0.0849	**0.02528**$^{\Rightarrow}$	**0.49001**$^{\Rightarrow}$	**0.02500**	**0.02166**	**0.02590**$^{\uparrow}$	**0.02679**$^{\rightarrow}$
	RP	**0.04458**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.77202**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25166**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19510**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22193**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03366**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.62804**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09170**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08951**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04458**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.77202**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25166**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19510**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22193**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03366**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.62804**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09170**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08951**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	CORI	0.02936	0.56435	0.139	0.123	0.0876	0.08336	0.02174	0.38914	0.022	0.024	0.0233	0.02115
	Taily	0.02895	0.55308	0.144	0.12333	0.0878	0.08648	0.02503	0.49994	0.021	0.01967	0.0238	0.02610
	R	**0.03022**$^{\rightarrow\Uparrow}$	**0.56897**	**0.14600**	**0.13700**$^{\Rightarrow\Uparrow}$	**0.09260**$^{\Rightarrow\Uparrow}$	**0.08625**	**0.02532**$^{\Rightarrow}$	**0.50411**$^{\Rightarrow}$	**0.02300**	**0.02000**$^{\leftarrow}$	**0.02490**	**0.02684**$^{\Rightarrow}$
	RT	**0.03022**$^{\rightarrow\Uparrow}$	**0.56897**$^{\Rightarrow}$	**0.14600**	**0.13700**$^{\Rightarrow\Uparrow}$	**0.09260**$^{\Rightarrow\Uparrow}$	**0.08625**	**0.02532**$^{\Rightarrow}$	**0.50411**$^{\Rightarrow}$	**0.02300**	**0.02000**$^{\leftarrow}$	**0.02490**	**0.02684**$^{\Rightarrow}$
	RP	**0.04545**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.79976**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25266**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19660**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22626**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03397**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.64168**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11233**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08970**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08913**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04545**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.79976**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.24100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25266**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19660**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22626**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03397**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.64168**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11233**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08970**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08913**$^{\Rightarrow\Uparrow\blacktriangle}$

Table B.2: Retrieval effectiveness on File Sharing environments (Scenario 1)

ASIS*		ASISWOR test-bed						ASISWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	CORI	0.02642	0.48121	0.167	0.12333	0.0798	0.0758	0.02143	0.41535	0.014	0.01833	0.0201	0.01994
	Taily	0.02608	0.46309	0.16	0.11966	0.0783	0.07104	0.01946	0.37962	0.019	0.01966	0.0218	0.02040
	R	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48191**$^{\uparrow}$	**0.17000**	**0.12900**$^{\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07851**$^{\rightarrow\Uparrow}$	**0.02219**$^{\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\uparrow}$	**0.02400**$^{\Rightarrow\uparrow}$	**0.02215**
	RT	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48191**$^{\uparrow}$	**0.17000**	**0.12900**$^{\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07851**$^{\rightarrow\Uparrow}$	**0.02219**$^{\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\uparrow}$	**0.02400**$^{\Rightarrow\uparrow}$	**0.02215**
	RP	**0.04343**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70714**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18533**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02818**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50008**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06067**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04343**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70714**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18533**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02818**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50008**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06067**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	CORI	0.02642	0.4792	0.164	0.12066	0.079	0.0744	0.02002	0.39688	0.014	0.018	0.0191	0.01869
	Taily	0.02574	0.46509	0.159	0.11866	0.0777	0.07013	0.0185	0.36757	0.016	0.01867	0.0211	0.01948
	R	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02219**$^{\rightarrow\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02400**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RT	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02219**$^{\rightarrow\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02400**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RP	**0.04344**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50023**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06750**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06068**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04344**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50023**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06750**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06068**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	CORI	0.02625	0.47761	0.163	0.12	0.0786	0.07327	0.01898	0.38319	0.014	0.01733	0.0188	0.01806
	Taily	0.02556	0.4644	0.158	0.119	0.0776	0.06978	0.01788	0.36117	0.015	0.019	0.0207	0.01876
	R	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\Rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\uparrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RT	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\Rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\uparrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RP	**0.04344**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50023**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06067**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04344**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50023**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06067**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	CORI	0.02601	0.47464	0.162	0.12	0.078	0.07247	0.01807	0.37005	0.014	0.01733	0.0187	0.01769
	Taily	0.02556	0.46792	0.16	0.119	0.0776	0.07035	0.01726	0.35735	0.014	0.01933	0.0204	0.01832
	R	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48210**$^{\uparrow}$	**0.17000**$^{\Rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\Uparrow}$	**0.02667**$^{\Rightarrow\uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RT	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48210**$^{\uparrow}$	**0.17000**$^{\Rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\Uparrow}$	**0.02667**$^{\Rightarrow\uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RP	**0.04344**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50023**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06067**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04344**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**$^{\rightarrow\Uparrow\triangle}$	**0.23900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.50023**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06067**$^{\Rightarrow\Uparrow\blacktriangle}$

Table B.3: Retrieval effectiveness on Uniformly Distributed environments (Scenario 1)

U*		UWOR test-bed						UWR test-bed					
%	Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
20%	CORI	0.02964	0.53141	0.19800	0.14400	0.09910	0.10867	0.02707	0.49861	0.01300	0.01700	0.02540	0.02691
	Taily	0.02862	0.50309	0.19700	0.14499	0.09710	0.10476	0.02467	0.45741	0.01400	0.01767	0.02580	0.02504
	R	**0.04052**$^{\Rightarrow\Uparrow}$	**0.57858**$^{\Rightarrow\Uparrow}$	**0.29800**$^{\Rightarrow\Uparrow}$	**0.21533**$^{\Rightarrow\Uparrow}$	**0.13620**$^{\Rightarrow\Uparrow}$	**0.16364**$^{\Rightarrow\Uparrow}$	**0.03498**$^{\Rightarrow\Uparrow}$	**0.49635**$^{\uparrow}$	**0.02900**$^{\Rightarrow\Uparrow}$	**0.03700**$^{\Rightarrow\Uparrow}$	**0.03750**$^{\Rightarrow\Uparrow}$	**0.03764**$^{\Rightarrow\Uparrow}$
	RT	**0.04053**$^{\Rightarrow\Uparrow}$	**0.57903**$^{\Rightarrow\Uparrow}$	**0.29800**$^{\Rightarrow\Uparrow}$	**0.21567**$^{\Rightarrow\Uparrow}$	**0.13630**$^{\Rightarrow\Uparrow}$	**0.16408**$^{\Rightarrow\Uparrow}$	**0.03498**$^{\Rightarrow\Uparrow}$	**0.49635**$^{\uparrow}$	**0.02900**$^{\Rightarrow\Uparrow}$	**0.03700**$^{\Rightarrow\Uparrow}$	**0.03750**$^{\Rightarrow\Uparrow}$	**0.03764**$^{\Rightarrow\Uparrow}$
	RP	**0.04857**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70675**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**	**0.25067**$^{\Rightarrow\Uparrow}$	**0.18820**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19417**$^{\Rightarrow\Uparrow}$	**0.03853**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55564**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10733**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08680**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08031**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04858**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70720**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**	**0.25067**$^{\Rightarrow\Uparrow}$	**0.18830**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19424**$^{\Rightarrow\Uparrow}$	**0.03853**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55564**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10733**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08680**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08031**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	CORI	0.02971	0.53442	0.18800	0.14267	0.09750	0.10947	0.02495	0.48878	0.01200	0.01600	0.02150	0.02398
	Taily	0.02860	0.51461	0.18800	0.14000	0.09570	0.10377	0.02347	0.44768	0.01300	0.01633	0.02380	0.02323
	R	**0.04151**$^{\Rightarrow\Uparrow}$	**0.61774**$^{\Rightarrow\Uparrow}$	**0.29100**$^{\Rightarrow\Uparrow}$	**0.21500**$^{\Rightarrow\Uparrow}$	**0.13820**$^{\Rightarrow\Uparrow}$	**0.16489**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03813**$^{\Rightarrow\Uparrow}$
	RT	**0.04151**$^{\Rightarrow\Uparrow}$	**0.61774**$^{\Rightarrow\Uparrow}$	**0.29100**$^{\Rightarrow\Uparrow}$	**0.21500**$^{\Rightarrow\Uparrow}$	**0.13820**$^{\Rightarrow\Uparrow}$	**0.16489**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03813**$^{\Rightarrow\Uparrow}$
	RP	**0.05099**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.76662**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**$^{\rightarrow\uparrow}$	**0.25300**$^{\Rightarrow\Uparrow\triangle}$	**0.19390**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20689**$^{\Rightarrow\Uparrow\triangle}$	**0.03860**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08660**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08068**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05099**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.76662**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**$^{\rightarrow\uparrow}$	**0.25300**$^{\Rightarrow\Uparrow\triangle}$	**0.19390**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20689**$^{\Rightarrow\Uparrow\triangle}$	**0.03860**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08660**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08068**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	CORI	0.02966	0.53270	0.19200	0.13800	0.09480	0.10784	0.02326	0.45085	0.01200	0.01567	0.02070	0.02210
	Taily	0.02855	0.51672	0.18800	0.13700	0.09290	0.10322	0.02229	0.42539	0.01400	0.01600	0.02180	0.02184
	R	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RT	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RP	**0.05108**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.77127**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**$^{\rightarrow\uparrow}$	**0.25334**$^{\Rightarrow\Uparrow\triangle}$	**0.19430**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20821**$^{\Rightarrow\Uparrow\triangle}$	**0.03860**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08650**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08067**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05108**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.77127**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**$^{\rightarrow\uparrow}$	**0.25334**$^{\Rightarrow\Uparrow\triangle}$	**0.19430**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20821**$^{\Rightarrow\Uparrow\triangle}$	**0.03860**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08650**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08067**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	CORI	0.02939	0.52887	0.19300	0.13833	0.09430	0.10808	0.02166	0.42868	0.01200	0.01533	0.02010	0.02094
	Taily	0.02833	0.51341	0.19100	0.13800	0.09220	0.1028	0.02165	0.42222	0.01200	0.01600	0.02060	0.02095
	R	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RT	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RP	**0.05108**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.77127**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**$^{\rightarrow\uparrow}$	**0.25334**$^{\Rightarrow\Uparrow\triangle}$	**0.19430**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20821**$^{\Rightarrow\Uparrow\triangle}$	**0.03860**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08650**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08067**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05108**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.77127**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25300**$^{\rightarrow\uparrow}$	**0.25334**$^{\Rightarrow\Uparrow\triangle}$	**0.19430**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20821**$^{\Rightarrow\Uparrow\triangle}$	**0.03860**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56655**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08650**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08067**$^{\Rightarrow\Uparrow\blacktriangle}$

Table B.4: Retrieval effectiveness on Digital Library environments (Scenario 2)

DL*		DLWOR test-bed						DLWR test-bed					
%	Method	Precision	Recall	P@10	P@30	P@100	MAP	Precision	Recall	P@10	P@30	P@100	MAP
20%	CORI	0.02888	0.53648	0.16900	0.13766	0.09150	0.09311	0.01733	0.22004	0.03100	0.02700	0.02760	0.01587
	Taily	0.02849	0.53061	0.16300	0.13000	0.08960	0.09006	0.02580	0.50380	0.02300	0.02133	0.02710	0.02872
	R	0.02735	0.48336$^{\Leftarrow\Downarrow}$	0.11800$^{\Leftarrow\Downarrow}$	0.12066$^{\leftarrow\downarrow}$	0.08520$^{\downarrow}$	0.07170$^{\Leftarrow\Downarrow}$	**0.02418**$^{\Rightarrow\downarrow}$	**0.44668**$^{\Rightarrow\Downarrow}$	**0.02900**$^{\uparrow}$	**0.02500**	0.02610	**0.02430**$^{\Rightarrow\Downarrow}$
	RT	0.02828	0.50166$^{\downarrow}$	0.11800$^{\Leftarrow\Downarrow}$	0.12267$^{\leftarrow}$	0.08670	0.07433$^{\Leftarrow\Downarrow}$	**0.02575**$^{\Rightarrow}$	**0.48813**$^{\Rightarrow}$	**0.02700**	**0.02466**	**0.02770**	**0.02721**$^{\Rightarrow}$
	RP	**0.03878**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.64610**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22500**$^{\uparrow\blacktriangle}$	**0.22967**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17727**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02915**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.51688**$^{\Rightarrow\blacktriangle}$	**0.12200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10966**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08610**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07332**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04312**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.71150**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.28400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.28500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20260**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23551**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03610**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.63776**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17967**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14460**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14692**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	CORI	0.02937	0.55104	0.15100	0.12733	0.09010	0.08714	0.01978	0.30967	0.02500	0.02533	0.02440	0.01974
	Taily	0.02909	0.54828	0.15700	0.12633	0.09020	0.08729	0.02564	0.50108	0.02200	0.02000	0.02510	0.02772
	R	0.02915	0.52427$^{\leftarrow\downarrow}$	0.12200$^{\Leftarrow\Downarrow}$	0.12400	0.08920	0.07795$^{\Leftarrow\Downarrow}$	**0.02533**$^{\Rightarrow}$	**0.48307**$^{\Rightarrow}$	**0.02700**$^{\uparrow}$	**0.02300**$^{\uparrow}$	**0.02590**	**0.02578**$^{\Rightarrow\downarrow}$
	RT	**0.02968**	0.54172	0.12200$^{\Leftarrow\Downarrow}$	0.12600	**0.09010**	0.07990$^{\leftarrow\downarrow}$	**0.02635**$^{\Rightarrow}$	**0.50571**$^{\Rightarrow}$	**0.02600**$^{\uparrow}$	**0.02333**$^{\uparrow}$	**0.02690**	**0.02776**$^{\Rightarrow}$
	RP	**0.04222**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.72109**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23533**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17970**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19522**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03134**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.57710**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11300**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10766**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08670**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07790**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04645**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.78327**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.28600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29033**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.21480**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.26318**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03831**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.68640**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17367**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14390**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15043**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	CORI	0.02972	0.56788	0.15700	0.12633	0.08960	0.08718	0.02125	0.36810	0.02600	0.02600	0.02430	0.02171
	Taily	0.02904	0.55292	0.14600	0.12267	0.08840	0.08732	0.02529	0.49908	0.02200	0.02033	0.02420	0.02673
	R	**0.02968**	0.54724$^{\leftarrow}$	0.11500$^{\Leftarrow\Downarrow}$	0.12366	**0.09000**	0.07920$^{\Leftarrow\downarrow}$	**0.02575**$^{\Rightarrow}$	**0.49812**$^{\Rightarrow}$	**0.02500**	**0.02233**	**0.02580**$^{\uparrow}$	**0.02653**$^{\rightarrow}$
	RT	**0.02993**$^{\uparrow}$	**0.55529**	0.11600$^{\Leftarrow\Downarrow}$	**0.12500**	**0.09060**$^{\uparrow}$	0.07992$^{\Leftarrow\downarrow}$	**0.02615**$^{\Rightarrow\uparrow}$	**0.51061**$^{\Rightarrow}$	**0.02400**	**0.02266**	**0.02640**$^{\uparrow}$	**0.02763**$^{\Rightarrow}$
	RP	**0.04377**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75899**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22700**$^{\rightarrow\Uparrow\blacktriangle}$	**0.23766**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18450**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20512**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03254**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.60648**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10500**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07871**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04783**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.81716**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.28600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29066**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.27281**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03905**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.70600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14190**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15079**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	CORI	0.02936	0.56435	0.13900	0.12300	0.08760	0.08336	0.02174	0.38914	0.02200	0.02400	0.02330	0.02115
	Taily	0.02895	0.55308	0.14400	0.12333	0.08780	0.08648	0.02503	0.49994	0.02100	0.01967	0.02380	0.02610
	R	**0.03009**$^{\Uparrow}$	**0.56617**	0.11600$^{\leftarrow\Downarrow}$	0.12267	**0.09000**$^{\rightarrow\uparrow}$	**0.08077**$^{\uparrow}$	**0.02588**$^{\Rightarrow\Uparrow}$	**0.50839**$^{\Rightarrow\Uparrow}$	**0.02300**	**0.02067**	**0.02440**	**0.02678**$^{\Rightarrow}$
	RT	**0.03024**$^{\rightarrow\Uparrow}$	**0.57382**$^{\Uparrow}$	0.11700$^{\leftarrow\Downarrow}$	**0.12333**	**0.09040**$^{\Rightarrow\Uparrow}$	0.08138	**0.02589**$^{\Rightarrow\Uparrow}$	**0.50896**$^{\Rightarrow\uparrow}$	**0.02300**	**0.02067**	**0.02440**	**0.02682**$^{\Rightarrow}$
	RP	**0.04459**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.78585**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23766**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18610**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20885**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03287**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.62084**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08280**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07837**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04868**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.84544**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.28600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22130**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.27728**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03910**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.71020**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15900**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16433**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14797**$^{\Rightarrow\Uparrow\blacktriangle}$

Table B.5: Reputation-based Retrieval effectiveness on File Sharing environments (Scenario 2)

ASIS*		ASISWOR test-bed						ASISWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	CORI	0.02642	0.48121	0.16700	0.12333	0.07980	0.0758	0.02143	0.41535	0.01400	0.01833	0.02010	0.01994
	Taily	0.02608	0.46309	0.16000	0.11966	0.07830	0.07104	0.01946	0.37962	0.01900	0.01966	0.02180	0.02040
	R	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48191**$^{\uparrow}$	**0.17000**	**0.12900**$^{\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07851**$^{\rightarrow\Uparrow}$	**0.02219**$^{\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\uparrow}$	**0.02400**$^{\Rightarrow\uparrow}$	**0.02215**
	RT	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48191**$^{\uparrow}$	**0.17000**	**0.12900**$^{\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07851**$^{\rightarrow\Uparrow}$	**0.02219**$^{\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\uparrow}$	**0.02400**$^{\Rightarrow\uparrow}$	**0.02214**
	RP	**0.04252**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.69360**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22500**$^{\uparrow}$	**0.22433**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16890**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16926**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02745**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.48852**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08300**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06060**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05242**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04619**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75096**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.27300**$^{\Rightarrow\Uparrow\triangle}$	**0.26567**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22530**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03282**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56247**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10610**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09899**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	CORI	0.02642	0.47920	0.16400	0.12066	0.07900	0.0744	0.02002	0.39688	0.01400	0.01800	0.01910	0.01869
	Taily	0.02574	0.46509	0.15900	0.11866	0.07770	0.07013	0.01850	0.36757	0.01600	0.01867	0.02110	0.01948
	R	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02219**$^{\rightarrow\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02400**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RT	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02219**$^{\rightarrow\Uparrow}$	**0.39545**	**0.02200**$^{\Rightarrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02400**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RP	**0.04253**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.69385**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22500**$^{\uparrow\rightarrow}$	**0.22433**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16890**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16927**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02747**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.48895**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08300**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06070**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05243**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04621**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75138**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.27300**$^{\Rightarrow\Uparrow\triangle}$	**0.26567**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22534**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03281**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56240**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10620**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09896**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	CORI	0.02625	0.47761	0.16300	0.12000	0.07860	0.07327	0.01898	0.38319	0.01400	0.01733	0.01880	0.01806
	Taily	0.02556	0.46440	0.15800	0.11900	0.07760	0.06978	0.01788	0.36117	0.01500	0.01900	0.02070	0.01876
	R	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\uparrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RT	**0.02908**$^{\Rightarrow\Uparrow}$	**0.48217**$^{\uparrow}$	**0.17000**$^{\rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\uparrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RP	**0.04253**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.69385**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22500**$^{\uparrow\rightarrow}$	**0.22433**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16890**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16927**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02747**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.48895**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08300**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06060**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05243**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04621**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75138**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.27300**$^{\Rightarrow\Uparrow\triangle}$	**0.26567**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22534**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03283**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56275**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12967**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10620**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09895**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	CORI	0.02601	0.47464	0.16200	0.12000	0.07800	0.07247	0.01807	0.37005	0.01400	0.01733	0.01870	0.01769
	Taily	0.02556	0.46792	0.16000	0.11900	0.07760	0.07035	0.01726	0.35735	0.01400	0.01933	0.02040	0.01832
	R	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48210**$^{\uparrow}$	**0.17000**$^{\rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\Uparrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RT	**0.02907**$^{\Rightarrow\Uparrow}$	**0.48210**$^{\uparrow}$	**0.17000**$^{\rightarrow\Uparrow}$	**0.12900**$^{\Rightarrow\Uparrow}$	**0.08420**$^{\Rightarrow\Uparrow}$	**0.07853**$^{\Rightarrow\Uparrow}$	**0.02218**$^{\Rightarrow\Uparrow}$	**0.39542**$^{\uparrow}$	**0.02200**$^{\Rightarrow\Uparrow}$	**0.02667**$^{\Rightarrow\Uparrow}$	**0.02390**$^{\Rightarrow\Uparrow}$	**0.02214**$^{\Rightarrow\Uparrow}$
	RP	**0.04253**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.69385**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22500**$^{\uparrow\rightarrow}$	**0.22433**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16890**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16927**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02747**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.48895**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08300**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.06060**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.05243**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04621**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75138**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.27300**$^{\Rightarrow\Uparrow\triangle}$	**0.26567**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22534**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03283**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.56275**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12967**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10620**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09895**$^{\Rightarrow\Uparrow\blacktriangle}$

Table B.6: Reputation-based Retrieval effectiveness on Uniformly Distributed environments (Scenario 2)

U*		UWOR test-bed						UWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	CORI	0.02964	0.53141	0.19800	0.14400	0.09910	0.10867	0.02707	0.49861	0.01300	0.01700	0.02540	0.02691
	Taily	0.02862	0.50309	0.19700	0.14499	0.09710	0.10476	0.02467	0.45741	0.01400	0.01767	0.02580	0.02504
	R	**0.03985**$^{\Rightarrow\Uparrow}$	**0.56047**$^{\Rightarrow\Uparrow}$	**0.29600**$^{\Rightarrow\Uparrow}$	**0.21166**$^{\Rightarrow\Uparrow}$	**0.13360**$^{\Rightarrow\Uparrow}$	**0.15964**$^{\Rightarrow\Uparrow}$	**0.03467**$^{\Rightarrow\Uparrow}$	**0.48008**	**0.02900**$^{\Rightarrow\Uparrow}$	**0.03700**$^{\Rightarrow\Uparrow}$	**0.03700**$^{\Rightarrow\Uparrow}$	**0.03696**$^{\Rightarrow\Uparrow}$
	RT	**0.03987**$^{\Rightarrow\Uparrow}$	**0.56138**$^{\Rightarrow\Uparrow}$	**0.29600**$^{\Rightarrow\Uparrow}$	**0.21233**$^{\Rightarrow\Uparrow}$	**0.13380**$^{\Rightarrow\Uparrow}$	**0.16051**$^{\Rightarrow\Uparrow}$	**0.03467**$^{\Rightarrow\Uparrow}$	**0.48008**	**0.02900**$^{\Rightarrow\Uparrow}$	**0.03700**$^{\Rightarrow\Uparrow}$	**0.03700**$^{\Rightarrow\Uparrow}$	**0.03696**$^{\Rightarrow\Uparrow}$
	RP	**0.04732**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.67506**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**	**0.22766**$^{\Rightarrow\Uparrow}$	**0.17460**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17332**$^{\Rightarrow\Uparrow}$	**0.03772**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.53016**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09733**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07205**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04734**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.67614**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23700**	**0.22766**$^{\Rightarrow\Uparrow}$	**0.17470**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17353**$^{\Rightarrow\Uparrow}$	**0.03772**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.53016**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09733**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07205**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	CORI	0.02971	0.53442	0.18800	0.14267	0.09750	0.10947	0.02495	0.48878	0.01200	0.01600	0.02150	0.02398
	Taily	0.02860	0.51461	0.18800	0.14000	0.09570	0.10377	0.02347	0.44768	0.01300	0.01633	0.02380	0.02323
	R	**0.04140**$^{\Rightarrow\Uparrow}$	**0.61208**$^{\Rightarrow\Uparrow}$	**0.29100**$^{\Rightarrow\Uparrow}$	**0.21500**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16450**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03813**$^{\Rightarrow\Uparrow}$
	RT	**0.04140**$^{\Rightarrow\Uparrow}$	**0.61208**$^{\Rightarrow\Uparrow}$	**0.29100**$^{\Rightarrow\Uparrow}$	**0.21500**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16450**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03813**$^{\Rightarrow\Uparrow}$
	RP	**0.05020**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.74535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23800**	**0.23000**$^{\Rightarrow\Uparrow}$	**0.18070**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18808**$^{\Rightarrow\Uparrow}$	**0.03800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55371**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07295**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05020**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.74535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23800**	**0.23000**$^{\Rightarrow\Uparrow}$	**0.18070**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18808**$^{\Rightarrow\Uparrow}$	**0.03800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55371**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07295**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	CORI	0.02966	0.53270	0.19200	0.13800	0.09480	0.10784	0.02326	0.45085	0.01200	0.01567	0.02070	0.02210
	Taily	0.02855	0.51672	0.18800	0.13700	0.09290	0.10322	0.02229	0.42539	0.01400	0.01600	0.02180	0.02184
	R	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RT	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RP	**0.05036**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23800**	**0.23000**$^{\Rightarrow\Uparrow}$	**0.18120**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18891**$^{\Rightarrow\Uparrow}$	**0.03800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55371**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07295**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05036**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23800**	**0.23000**$^{\Rightarrow\Uparrow}$	**0.18120**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18891**$^{\Rightarrow\Uparrow}$	**0.03800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55371**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07295**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	CORI	0.02939	0.52887	0.19300	0.13833	0.09430	0.10808	0.02166	0.42868	0.01200	0.01533	0.02010	0.02094
	Taily	0.02833	0.51341	0.19100	0.13800	0.09220	0.1028	0.02165	0.42222	0.01200	0.01600	0.02060	0.02095
	R	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RT	**0.04151**$^{\Rightarrow\Uparrow}$	**0.62181**$^{\Rightarrow\Uparrow}$	**0.29200**$^{\Rightarrow\Uparrow}$	**0.21466**$^{\Rightarrow\Uparrow}$	**0.13800**$^{\Rightarrow\Uparrow}$	**0.16511**$^{\Rightarrow\Uparrow}$	**0.03501**$^{\Rightarrow\Uparrow}$	**0.50751**$^{\Rightarrow\Uparrow}$	**0.02700**$^{\Rightarrow\Uparrow}$	**0.03666**$^{\Rightarrow\Uparrow}$	**0.03790**$^{\Rightarrow\Uparrow}$	**0.03812**$^{\Rightarrow\Uparrow}$
	RP	**0.05036**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23800**	**0.23000**$^{\Rightarrow\Uparrow}$	**0.18120**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18891**$^{\Rightarrow\Uparrow}$	**0.03800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55371**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07295**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.05036**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.75520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.23800**	**0.23000**$^{\Rightarrow\Uparrow}$	**0.18120**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18891**$^{\Rightarrow\Uparrow}$	**0.03800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.55371**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.10600**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09700**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08010**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07295**$^{\Rightarrow\Uparrow\blacktriangle}$

B.2 Robustness in Varying of Training and Testing Boundaries

Table B.7: Retrieval effectiveness Boundaries DL Environment (25-75)%

DL*		DLWOR test-bed						DLWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.02868	0.53541	0.17067	0.12933	0.09253	0.09024	0.02544	0.50416	0.01333	0.01289	0.02147	0.02509
	CORI	0.02908	0.53199	0.17067	0.13733	0.09413	0.09259	0.01610	0.21225	0.02000	0.02000	0.02320	0.01377
	R/RP	0.02627	0.45132	0.10800	0.11244	0.08200	0.06573	**0.02278**$^{\Rightarrow}$	**0.39307**$^{\Rightarrow}$	**0.01733**$^{\Uparrow}$	**0.01422**$^{\Uparrow}$	**0.02320**$^{\uparrow}$	**0.02045**$^{\Rightarrow}$
	RT	**0.03136**$^{\rightarrow\Uparrow\blacktriangle}$	**0.58301**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14267**$^{\blacktriangle}$	**0.13911**$^{\uparrow\blacktriangle}$	**0.10080**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09011**$^{\blacktriangle}$	**0.02738**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.52959**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02000**$^{\Uparrow}$	**0.01778**$^{\Uparrow}$	**0.02613**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02919**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04276**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.73493**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.47200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.39022**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22853**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34606**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03535**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.64137**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.19511**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13640**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14969**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	Taily	0.02936	0.55385	0.16133	0.12489	0.09360	0.08655	0.02518	0.51160	0.01333	0.01200	0.02000	0.02427
	CORI	0.02953	0.54513	0.14933	0.12577	0.09253	0.08461	0.01933	0.30625	0.01467	0.01733	0.01907	0.01761
	R/RP	0.02839	0.49944	0.11600	0.12178	0.09000	0.07336	**0.02382**$^{\Rightarrow}$	**0.45994**$^{\Rightarrow}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.01511**$^{\Uparrow}$	**0.02013**	**0.02203**$^{\Rightarrow}$
	RT	**0.03093**$^{\blacktriangle}$	**0.57793**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12533**$^{\blacktriangle}$	**0.12800**	**0.09733**$^{\Rightarrow\Uparrow\blacktriangle}$	0.08392	**0.02552**$^{\Rightarrow\blacktriangle}$	**0.51386**$^{\Rightarrow\blacktriangle}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.01555**$^{\Uparrow}$	**0.02173**$^{\rightarrow}$	**0.02569**$^{\Rightarrow\triangle}$
	RPT	**0.04284**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.74087**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.47200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.38978**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22760**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34493**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03472**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.64644**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17733**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17422**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12427**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13660**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	Taily	0.02917	0.55434	0.14533	0.12000	0.09120	0.08496	0.02477	0.50602	0.01333	0.01244	0.01920	0.02316
	CORI	0.02988	0.56585	0.15600	0.12267	0.09213	0.08411	0.02074	0.34939	0.01600	0.01778	0.01880	0.01905
	R/RP	**0.02931**	0.53438	0.11067	0.12044	0.09120	0.07469	**0.02415**$^{\Rightarrow}$	**0.50001**$^{\Rightarrow}$	**0.01600**$^{\uparrow}$	**0.01333**$^{\uparrow}$	**0.01960**$^{\Rightarrow}$	**0.02277**$^{\Rightarrow}$
	RT	**0.03023**	**0.57426**$^{\Rightarrow\Uparrow\blacktriangle}$	0.11600	**0.12400**	**0.09480**	0.08036	**0.02476**$^{\Rightarrow}$	**0.51292**$^{\Rightarrow}$	**0.01600**$^{\uparrow}$	**0.01378**$^{\Uparrow}$	**0.02013**$^{\rightarrow}$	**0.02408**$^{\Rightarrow\triangle}$
	RPT	**0.04239**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.74091**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.47200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.38845**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22680**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34340**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03451**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.65151**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16133**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16266**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11987**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13064**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	Taily	0.02916	0.55464	0.14000	0.11955	0.09013	0.08358	0.02438	0.50571	0.01200	0.01200	0.01880	0.02247
	CORI	0.02956	0.56351	0.13333	0.11867	0.08960	0.07964	0.02105	0.38265	0.01467	0.01556	0.01760	0.01808
	R/RP	**0.02993**	**0.56323**	0.11200	**0.12222**$^{\Rightarrow\Uparrow}$	**0.09240**	0.07889	**0.02423**$^{\Rightarrow}$	**0.50585**$^{\Rightarrow}$	**0.01467**$^{\uparrow}$	**0.01244**	**0.01907**$^{\Rightarrow\uparrow}$	**0.02291**$^{\Rightarrow}$
	RT	**0.03020**	**0.57413**$^{\Rightarrow\Uparrow\blacktriangle}$	0.11200	**0.12266**$^{\Rightarrow\Uparrow}$	**0.09307**	**0.08001**$^{\triangle}$	**0.02421**$^{\Rightarrow}$	**0.50583**$^{\Rightarrow}$	**0.01467**$^{\uparrow}$	**0.01244**	**0.01907**$^{\Rightarrow\uparrow}$	**0.02290**$^{\Rightarrow}$
	RPT	**0.04244**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.74452**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.47200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.38800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22667**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34327**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03440**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.65225**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15333**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12586**$^{\Rightarrow\Uparrow\blacktriangle}$

Table B.8: Retrieval effectiveness Boundaries DL Environment (50-50)%

DL*		DLWOR test-bed						DLWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.02935	0.50663	0.16200	0.13133	0.09200	0.08369	0.02579	0.46825	0.00800	0.00800	0.01500	0.01986
	CORI	0.02945	0.49511	0.15200	0.13799	0.09120	0.07932	0.01534	0.19200	0.01000	0.01466	0.02080	0.00937
	R/RP	0.02700	0.42769	0.10200	0.11133	0.08480	0.05847	**0.02370**$^{\Rightarrow}$	**0.37170**$^{\Rightarrow}$	**0.01200**$^{\Uparrow}$	**0.00933**	**0.01700**$^{\Uparrow}$	**0.01778**$^{\Rightarrow}$
	RT	**0.03242**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.54122**$^{\Rightarrow\uparrow\blacktriangle}$	0.13800	**0.13733**$^{\blacktriangle}$	**0.10060**$^{\blacktriangle}$	**0.08138**$^{\Rightarrow\blacktriangle}$	**0.02810**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.48224**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.01400**$^{\rightarrow\Uparrow}$	**0.01000**$^{\Uparrow}$	**0.01840**$^{\Uparrow}$	**0.02391**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04492**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.71076**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.41867**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25040**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34286**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03778**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.60748**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.20800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14680**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14614**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	Taily	0.03013	0.51962	0.15000	0.12467	0.09280	0.07803	0.02569	0.47002	0.00800	0.00800	0.01420	0.01880
	CORI	0.03041	0.51255	0.14400	0.12533	0.09080	0.07492	0.01863	0.26266	0.00600	0.01066	0.01220	0.01071
	R/RP	0.02939	0.46111	0.10400	0.11667	0.09060	0.06431	**0.02480**$^{\Rightarrow}$	**0.42518**$^{\Rightarrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.00933**	**0.01380**$^{\Rightarrow}$	**0.01807**$^{\Rightarrow}$
	RT	**0.03192**$^{\rightarrow\Uparrow\triangle}$	**0.53841**$^{\Rightarrow\Uparrow\blacktriangle}$	0.11600	0.12333	**0.09740**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.07464**$^{\blacktriangle}$	**0.02638**$^{\Rightarrow\uparrow\triangle}$	**0.47189**$^{\Rightarrow\blacktriangle}$	**0.01200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.00933**	**0.01560**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.02046**$^{\Rightarrow\Uparrow\blacktriangle}$
	RPT	**0.04522**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.71888**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.41934**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25020**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34254**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03722**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.61063**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.17400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.18333**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13520**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13252**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	Taily	0.02997	0.51619	0.14000	0.11800	0.09000	0.07637	0.02529	0.46261	0.00600	0.00800	0.01340	0.01797
	CORI	0.03059	0.52648	0.15000	0.11867	0.09000	0.07537	0.02082	0.30353	0.00800	0.01133	0.01220	0.01287
	R/RP	**0.03030**	0.49312	0.10000	0.11733	**0.09040**	0.06478	**0.02509**$^{\Rightarrow}$	**0.46398**$^{\Rightarrow}$	**0.01000**$^{\rightarrow\Uparrow}$	**0.00800**	**0.01440**$^{\Rightarrow\Uparrow}$	**0.01859**$^{\Rightarrow\uparrow}$
	RT	**0.03118**$^{\Uparrow}$	**0.53207**$^{\uparrow\blacktriangle}$	0.10800	**0.12200**	**0.09380**$^{\uparrow}$	**0.07112**$^{\blacktriangle}$	**0.02561**$^{\Rightarrow}$	**0.47458**$^{\Rightarrow\uparrow}$	**0.01000**$^{\Rightarrow\Uparrow}$	**0.00800**	**0.01460**$^{\Rightarrow\Uparrow}$	**0.01918**$^{\Rightarrow\Uparrow\triangle}$
	RPT	**0.04482**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.71765**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.41800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25020**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34103**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03701**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.61758**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.16800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.13040**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12622**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	Taily	0.03001	0.51660	0.13600	0.11933	0.08900	0.07493	0.02493	0.45848	0.00600	0.00800	0.01340	0.01735
	CORI	0.03039	0.52404	0.13200	0.11601	0.08740	0.07079	0.02128	0.32107	0.00800	0.00933	0.01100	0.01240
	R/RP	**0.03088**	**0.51699**	0.10400	**0.12067**$^{\rightarrow}$	**0.09140**$^{\Rightarrow\uparrow}$	0.06995	**0.02485**$^{\Rightarrow}$	**0.46441**$^{\Rightarrow\uparrow}$	**0.00800**$^{\uparrow}$	**0.00800**	**0.01360**$^{\rightarrow}$	**0.01787**$^{\Rightarrow}$
	RT	**0.03115**$^{\Uparrow}$	**0.53070**$^{\uparrow}$	0.10400	**0.12133**$^{\rightarrow}$	**0.09200**$^{\Rightarrow\uparrow}$	**0.07088**$^{\Rightarrow}$	**0.02485**$^{\Rightarrow}$	**0.46441**$^{\Rightarrow\uparrow}$	**0.00800**$^{\uparrow}$	**0.00800**	**0.01360**$^{\rightarrow}$	**0.01786**$^{\Rightarrow}$
	RPT	**0.04493**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.72035**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.41800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.25040**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.34100**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03695**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.61724**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.14000**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.15466**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12400**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.12001**$^{\Rightarrow\Uparrow\blacktriangle}$

B.2 Robustness in Varying of Training and Testing Boundaries

Table B.9: Retrieval effectiveness Boundaries DL Environment (75-25)%

DL*		DLWOR test-bed						DLWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.03559	0.51058	0.16400	0.15200	0.10960	0.10978	0.03007	0.45294	0.01200	0.01333	0.02000	0.02364
	CORI	0.03535	0.48740	0.16000	0.15199	0.10240	0.09878	0.01932	0.18255	0.01600	0.01866	0.01560	0.01064
	R/RP	0.03303	0.44391	0.13600	0.13600	**0.10440**	0.08090	**0.02913**$^{\Rightarrow}$	**0.38166**$^{\Rightarrow}$	**0.02000**$^{\Rightarrow\uparrow}$	**0.01466**$^{\uparrow}$	**0.02000**$^{\Rightarrow}$	**0.02350**$^{\Rightarrow}$
	RT	**0.03884**$^{\Rightarrow\uparrow\blacktriangle}$	**0.55834**$^{\Rightarrow\uparrow\blacktriangle}$	**0.17600**$^{\Rightarrow\uparrow\blacktriangle}$	**0.16267**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.10273**$^{\Rightarrow\blacktriangle}$	**0.03292**$^{\Rightarrow\triangle}$	**0.48647**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02400**$^{\Rightarrow\uparrow\blacktriangle}$	**0.01600**$^{\uparrow\blacktriangle}$	**0.02040**$^{\Rightarrow}$	**0.02844**$^{\Rightarrow\uparrow\blacktriangle}$
	RPT	**0.05156**$^{\Rightarrow\uparrow\blacktriangle}$	**0.72994**$^{\Rightarrow\uparrow\blacktriangle}$	**0.45200**$^{\Rightarrow\uparrow\blacktriangle}$	**0.38667**$^{\Rightarrow\uparrow\blacktriangle}$	**0.23920**$^{\Rightarrow\uparrow\blacktriangle}$	**0.35496**$^{\Rightarrow\uparrow\blacktriangle}$	**0.04336**$^{\Rightarrow\uparrow\blacktriangle}$	**0.61709**$^{\Rightarrow\uparrow\blacktriangle}$	**0.20400**$^{\Rightarrow\uparrow\blacktriangle}$	**0.20000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.14080**$^{\Rightarrow\uparrow\blacktriangle}$	**0.16998**$^{\Rightarrow\uparrow\blacktriangle}$
30%	Taily	0.03691	0.53358	0.18400	0.15600	0.11360	0.10334	0.03043	0.47879	0.01200	0.01333	0.01880	0.02263
	CORI	0.03751	0.51905	0.18400	0.15599	0.11160	0.10011	0.02262	0.23726	0.01200	0.01600	0.01400	0.01297
	R/RP	0.03649	0.47890	0.12800	0.13867	0.11240	0.08886	**0.02984**$^{\Rightarrow}$	**0.44250**$^{\Rightarrow}$	**0.02000**$^{\Rightarrow\uparrow}$	**0.01600**$^{\uparrow}$	**0.01680**$^{\Rightarrow}$	**0.02240**$^{\Rightarrow}$
	RT	**0.03889**$^{\Rightarrow\uparrow\blacktriangle}$	**0.54931**$^{\Rightarrow\uparrow\blacktriangle}$	0.13600	0.14133	**0.11760**	0.09596	**0.03129**$^{\Rightarrow}$	**0.49533**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.01600**$^{\uparrow}$	**0.01840**$^{\Rightarrow\blacktriangle}$	**0.02454**$^{\Rightarrow\uparrow\blacktriangle}$
	RPT	**0.05233**$^{\Rightarrow\uparrow\blacktriangle}$	**0.73068**$^{\Rightarrow\uparrow\blacktriangle}$	**0.45200**$^{\Rightarrow\uparrow\blacktriangle}$	**0.38800**$^{\Rightarrow\uparrow\blacktriangle}$	**0.23960**$^{\Rightarrow\uparrow\blacktriangle}$	**0.35576**$^{\Rightarrow\uparrow\blacktriangle}$	**0.04261**$^{\Rightarrow\uparrow\blacktriangle}$	**0.62368**$^{\Rightarrow\uparrow\blacktriangle}$	**0.16400**$^{\Rightarrow\uparrow\blacktriangle}$	**0.18666**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12760**$^{\Rightarrow\uparrow\blacktriangle}$	**0.15788**$^{\Rightarrow\uparrow\blacktriangle}$
40%	Taily	0.03719	0.53666	0.16400	0.14800	0.11120	0.10349	0.03011	0.47442	0.01200	0.01333	0.01760	0.02191
	CORI	0.03767	0.54272	0.18000	0.14667	0.11040	0.10109	0.02436	0.27171	0.01600	0.02133	0.01600	0.01536
	R/RP	**0.03768**	0.51416	0.12800	0.14133	**0.11120**	0.08887	**0.02950**$^{\Rightarrow}$	**0.47408**$^{\Rightarrow}$	**0.01600**$^{\uparrow}$	**0.01466**$^{\uparrow}$	**0.01720**$^{\Rightarrow}$	**0.02206**$^{\Rightarrow}$
	RT	**0.03841**$^{\rightarrow\uparrow\triangle}$	**0.54960**$^{\uparrow\blacktriangle}$	0.13200	0.14534	**0.11360**	0.09469	**0.02994**$^{\Rightarrow}$	**0.48204**$^{\Rightarrow\uparrow\triangle}$	**0.01600**$^{\uparrow}$	**0.01466**$^{\uparrow}$	**0.01760**$^{\Rightarrow}$	**0.02260**$^{\Rightarrow}$
	RPT	**0.05205**$^{\Rightarrow\uparrow\blacktriangle}$	**0.73467**$^{\Rightarrow\uparrow\blacktriangle}$	**0.45200**$^{\Rightarrow\uparrow\blacktriangle}$	**0.38667**$^{\Rightarrow\uparrow\blacktriangle}$	**0.24000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.35538**$^{\Rightarrow\uparrow\blacktriangle}$	**0.04194**$^{\Rightarrow\uparrow\blacktriangle}$	**0.60947**$^{\Rightarrow\uparrow\blacktriangle}$	**0.13600**$^{\Rightarrow\uparrow\blacktriangle}$	**0.17333**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12240**$^{\Rightarrow\uparrow\blacktriangle}$	**0.15114**$^{\Rightarrow\uparrow\blacktriangle}$
50%	Taily	0.03731	0.53748	0.15600	0.14933	0.11040	0.10174	0.02975	0.47166	0.01200	0.01333	0.01760	0.02130
	CORI	0.03755	0.54245	0.15600	0.14534	0.10720	0.09452	0.02503	0.28934	0.01600	0.01733	0.01480	0.01450
	R/RP	**0.03848**$^{\Rightarrow\uparrow\blacktriangle}$	0.52832	0.13200	0.14400	**0.11200**	0.09374	**0.02907**$^{\Rightarrow}$	**0.47059**$^{\Rightarrow}$	**0.01600**$^{\uparrow}$	**0.01466**$^{\uparrow}$	**0.01640**$^{\Rightarrow}$	**0.02117**$^{\Rightarrow}$
	RT	**0.03878**$^{\Rightarrow\uparrow}$	**0.55180**$^{\Rightarrow\uparrow\blacktriangle}$	0.13200	**0.14534**	**0.11240**	**0.09488**$^{\triangle}$	**0.02907**$^{\Rightarrow}$	**0.47059**$^{\Rightarrow}$	**0.01600**$^{\uparrow}$	**0.01466**$^{\uparrow}$	**0.01640**$^{\Rightarrow}$	**0.02116**$^{\Rightarrow}$
	RPT	**0.05230**$^{\Rightarrow\uparrow\blacktriangle}$	**0.73999**$^{\Rightarrow\uparrow\blacktriangle}$	**0.45200**$^{\Rightarrow\uparrow\blacktriangle}$	**0.38667**$^{\Rightarrow\uparrow\blacktriangle}$	**0.24040**$^{\Rightarrow\uparrow\blacktriangle}$	**0.35565**$^{\Rightarrow\uparrow\blacktriangle}$	**0.04195**$^{\Rightarrow\uparrow\blacktriangle}$	**0.61226**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12800**$^{\Rightarrow\uparrow\blacktriangle}$	**0.15600**$^{\Rightarrow\uparrow\blacktriangle}$	**0.11600**$^{\Rightarrow\uparrow\blacktriangle}$	**0.14534**$^{\Rightarrow\uparrow\blacktriangle}$

Table B.10: Retrieval effectiveness Boundaries ASIS Environment (25-75)%

ASIS*		ASISWOR test-bed						ASISWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.02621	0.45636	0.16000	0.12133	0.08267	0.06242	0.01817	0.37154	0.00933	0.01378	0.01747	0.01676
	CORI	0.02659	0.46873	0.16400	0.12355	0.08333	0.06646	0.02043	0.40155	0.00800	0.01289	0.01587	0.01684
	R/RP	**0.02877**$^{\Rightarrow\uparrow}$	**0.47087**$^{\Rightarrow\uparrow}$	**0.16800**$^{\rightarrow\uparrow}$	**0.13200**$^{\Rightarrow\uparrow}$	**0.08827**$^{\Rightarrow\uparrow}$	**0.06925**$^{\Rightarrow\uparrow}$	**0.02082**$^{\Rightarrow\uparrow}$	**0.37573**	**0.01600**$^{\Rightarrow\uparrow}$	**0.02000**$^{\Rightarrow\uparrow}$	**0.01867**$^{\Rightarrow\uparrow}$	**0.01895**$^{\Rightarrow\uparrow}$
	RT	**0.02878**$^{\Rightarrow\uparrow\blacktriangle}$	**0.47108**$^{\Rightarrow\uparrow\blacktriangle}$	**0.16800**$^{\rightarrow\uparrow\blacktriangle}$	**0.13200**$^{\Rightarrow\uparrow\blacktriangle}$	**0.08840**$^{\Rightarrow\uparrow\blacktriangle}$	**0.06930**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02082**$^{\Rightarrow\uparrow\blacktriangle}$	**0.37573**$^{\blacktriangle}$	**0.01600**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.01867**$^{\Rightarrow\uparrow\blacktriangle}$	**0.01895**$^{\Rightarrow\uparrow\blacktriangle}$
	RPT	**0.04109**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66514**$^{\Rightarrow\uparrow\blacktriangle}$	**0.44133**$^{\Rightarrow\uparrow\blacktriangle}$	**0.36623**$^{\Rightarrow\uparrow\blacktriangle}$	**0.21280**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29121**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02980**$^{\Rightarrow\uparrow\blacktriangle}$	**0.51813**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12667**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12311**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09493**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09173**$^{\Rightarrow\uparrow\blacktriangle}$
30%	Taily	0.02587	0.45975	0.15867	0.12089	0.08200	0.06153	0.01691	0.35370	0.00800	0.01289	0.01667	0.01579
	CORI	0.02664	0.46671	0.16133	0.12177	0.08293	0.06503	0.01875	0.38166	0.00800	0.01244	0.01480	0.01547
	R/RP/RT	**0.02880**$^{\Rightarrow\uparrow}$	**0.47143**$^{\Rightarrow\uparrow}$	**0.16800**$^{\uparrow}$	**0.13200**$^{\Rightarrow\uparrow}$	**0.08840**$^{\Rightarrow\uparrow}$	**0.06934**$^{\Rightarrow\uparrow}$	**0.02082**$^{\Rightarrow\uparrow}$	**0.37573**$^{\uparrow}$	**0.01600**$^{\Rightarrow\uparrow}$	**0.02000**$^{\Rightarrow\uparrow}$	**0.01867**$^{\Rightarrow\uparrow}$	**0.01894**$^{\Rightarrow\uparrow}$
	RPT	**0.04110**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66548**$^{\Rightarrow\uparrow\blacktriangle}$	**0.44133**$^{\Rightarrow\uparrow\blacktriangle}$	**0.36623**$^{\Rightarrow\uparrow\blacktriangle}$	**0.21280**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29123**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02981**$^{\Rightarrow\uparrow\blacktriangle}$	**0.51815**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12533**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12311**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09480**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09175**$^{\Rightarrow\uparrow\blacktriangle}$
40%	Taily	0.02572	0.45943	0.15733	0.12089	0.08200	0.06134	0.01617	0.34547	0.00800	0.01289	0.01613	0.01505
	CORI	0.02644	0.46556	0.15867	0.12089	0.08280	0.06371	0.01745	0.36525	0.00800	0.01200	0.01440	0.01474
	R/RP/RT	**0.02880**$^{\Rightarrow\uparrow}$	**0.47143**$^{\Rightarrow\uparrow}$	**0.16800**$^{\Rightarrow\uparrow}$	**0.13200**$^{\Rightarrow\uparrow}$	**0.08840**$^{\Rightarrow\uparrow}$	**0.06934**$^{\Rightarrow\uparrow}$	**0.02082**$^{\Rightarrow\uparrow}$	**0.37573**$^{\Rightarrow\uparrow}$	**0.01600**$^{\Rightarrow\uparrow}$	**0.02000**$^{\Rightarrow\uparrow}$	**0.01867**$^{\Rightarrow\uparrow}$	**0.01895**$^{\Rightarrow\uparrow}$
	RPT	**0.04110**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66548**$^{\Rightarrow\uparrow\blacktriangle}$	**0.44133**$^{\Rightarrow\uparrow\blacktriangle}$	**0.36623**$^{\Rightarrow\uparrow\blacktriangle}$	**0.21280**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29123**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02981**$^{\Rightarrow\uparrow\blacktriangle}$	**0.51815**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12533**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12311**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09480**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09176**$^{\Rightarrow\uparrow\blacktriangle}$
50%	Taily	0.02563	0.46036	0.15867	0.12044	0.08187	0.06172	0.01539	0.33793	0.00800	0.01333	0.01587	0.01465
	CORI	0.02623	0.46351	0.15867	0.12089	0.08200	0.06297	0.01640	0.35069	0.00800	0.01200	0.01427	0.01431
	R/RP/RT	**0.02880**$^{\Rightarrow\uparrow}$	**0.47143**$^{\Rightarrow\uparrow}$	**0.16800**$^{\Rightarrow\uparrow}$	**0.13200**$^{\Rightarrow\uparrow}$	**0.08840**$^{\Rightarrow\uparrow}$	**0.06934**$^{\Rightarrow\uparrow}$	**0.02081**$^{\Rightarrow\uparrow}$	**0.37570**$^{\Rightarrow\uparrow}$	**0.01600**$^{\Rightarrow\uparrow}$	**0.02000**$^{\Rightarrow\uparrow}$	**0.01853**$^{\Rightarrow\uparrow}$	**0.01894**$^{\Rightarrow\uparrow}$
	RPT	**0.04110**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66548**$^{\Rightarrow\uparrow\blacktriangle}$	**0.44133**$^{\Rightarrow\uparrow\blacktriangle}$	**0.36623**$^{\Rightarrow\uparrow\blacktriangle}$	**0.21280**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29123**$^{\Rightarrow\uparrow\blacktriangle}$	**0.02980**$^{\Rightarrow\uparrow\blacktriangle}$	**0.51813**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12533**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12311**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09480**$^{\Rightarrow\uparrow\blacktriangle}$	**0.09175**$^{\Rightarrow\uparrow\blacktriangle}$

Table B.11: Retrieval effectiveness Boundaries ASIS Environment (50-50)%

ASIS*		ASISWOR test-bed						ASISWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.02732	0.45062	0.15200	0.12133	0.08460	0.05738	0.01828	0.33869	0.00200	0.00866	0.01140	0.01137
	CORI	0.02790	0.45946	0.15400	0.12200	0.08460	0.06003	0.02144	0.37937	0.00200	0.00800	0.01100	0.01259
	R/RP/RT	**0.02960**$^{\Rightarrow\uparrow}$	**0.46957**$^{\Rightarrow\uparrow}$	**0.16800**$^{\Rightarrow\uparrow}$	**0.13267**$^{\Rightarrow\uparrow}$	**0.09100**$^{\Rightarrow\uparrow}$	**0.06533**$^{\Rightarrow\uparrow}$	**0.02051**$^{\uparrow}$	**0.34558**$^{\uparrow}$	**0.00600**$^{\Rightarrow\uparrow}$	**0.01200**$^{\Rightarrow\uparrow}$	**0.01280**$^{\Rightarrow\uparrow}$	**0.01254**$^{\uparrow}$
	RPT	**0.04328**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66244**$^{\Rightarrow\uparrow\blacktriangle}$	**0.47000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.40067**$^{\Rightarrow\uparrow\blacktriangle}$	**0.23640**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29819**$^{\Rightarrow\uparrow\blacktriangle}$	**0.03127**$^{\Rightarrow\uparrow\blacktriangle}$	**0.50853**$^{\Rightarrow\uparrow\blacktriangle}$	**0.11200**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12534**$^{\Rightarrow\uparrow\blacktriangle}$	**0.10280**$^{\Rightarrow\uparrow\blacktriangle}$	**0.08131**$^{\Rightarrow\uparrow\blacktriangle}$
30%	Taily	0.02690	0.44734	0.15200	0.12000	0.08340	0.05614	0.01662	0.31469	0.00200	0.00800	0.01080	0.01027
	CORI	0.02804	0.45695	0.15400	0.12000	0.08500	0.06027	0.01936	0.35387	0.00200	0.00733	0.00980	0.01105
	R/RP/RT	**0.02962**$^{\Rightarrow\uparrow}$	**0.47008**$^{\Rightarrow\uparrow}$	**0.16800**$^{\Rightarrow\uparrow}$	**0.13267**$^{\Rightarrow\uparrow}$	**0.09100**$^{\Rightarrow\uparrow}$	**0.06539**$^{\Rightarrow\uparrow}$	**0.02051**$^{\Rightarrow\uparrow}$	**0.34558**$^{\uparrow}$	**0.00600**$^{\Rightarrow\uparrow}$	**0.01200**$^{\Rightarrow\uparrow}$	**0.01280**$^{\Rightarrow\uparrow}$	**0.01253**$^{\Rightarrow\uparrow}$
	RPT	**0.04330**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66295**$^{\Rightarrow\uparrow\blacktriangle}$	**0.47000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.40067**$^{\Rightarrow\uparrow\blacktriangle}$	**0.23640**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29822**$^{\Rightarrow\uparrow\blacktriangle}$	**0.03127**$^{\Rightarrow\uparrow\blacktriangle}$	**0.50853**$^{\Rightarrow\uparrow\blacktriangle}$	**0.11000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12534**$^{\Rightarrow\uparrow\blacktriangle}$	**0.10260**$^{\Rightarrow\uparrow\blacktriangle}$	**0.08134**$^{\Rightarrow\uparrow\blacktriangle}$
40%	Taily	0.02678	0.44612	0.15400	0.11933	0.08340	0.05595	0.01584	0.30591	0.00200	0.00800	0.01020	0.00984
	CORI	0.02790	0.45743	0.15600	0.12000	0.08480	0.05978	0.01778	0.33592	0.00200	0.00667	0.00920	0.01014
	R/RP/RT	**0.02962**$^{\Rightarrow\uparrow}$	**0.47008**$^{\Rightarrow\uparrow}$	**0.16800**$^{\Rightarrow\uparrow}$	**0.13267**$^{\Rightarrow\uparrow}$	**0.09100**$^{\Rightarrow\uparrow}$	**0.06539**$^{\Rightarrow\uparrow}$	**0.02051**$^{\Rightarrow\uparrow}$	**0.34558**$^{\Rightarrow\uparrow}$	**0.00600**$^{\Rightarrow\uparrow}$	**0.01200**$^{\Rightarrow\uparrow}$	**0.01280**$^{\Rightarrow\uparrow}$	**0.01253**$^{\Rightarrow\uparrow}$
	RPT	**0.04330**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66295**$^{\Rightarrow\uparrow\blacktriangle}$	**0.47000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.40067**$^{\Rightarrow\uparrow\blacktriangle}$	**0.23640**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29822**$^{\Rightarrow\uparrow\blacktriangle}$	**0.03127**$^{\Rightarrow\uparrow\blacktriangle}$	**0.50853**$^{\Rightarrow\uparrow\blacktriangle}$	**0.11000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12534**$^{\Rightarrow\uparrow\blacktriangle}$	**0.10260**$^{\Rightarrow\uparrow\blacktriangle}$	**0.08134**$^{\Rightarrow\uparrow\blacktriangle}$
50%	Taily	0.02670	0.44780	0.15600	0.11933	0.08340	0.05689	0.01484	0.29858	0.00200	0.00800	0.00960	0.00940
	CORI	0.02768	0.45497	0.15600	0.12000	0.08400	0.05904	0.01626	0.31500	0.00200	0.00667	0.00920	0.00954
	R/RP/RT	**0.02962**$^{\Rightarrow\uparrow}$	**0.47008**$^{\Rightarrow\uparrow}$	**0.16800**$^{\Rightarrow\uparrow}$	**0.13267**$^{\Rightarrow\uparrow}$	**0.09100**$^{\Rightarrow\uparrow}$	**0.06539**$^{\Rightarrow\uparrow}$	**0.02049**$^{\Rightarrow\uparrow}$	**0.34553**$^{\Rightarrow\uparrow}$	**0.00600**$^{\Rightarrow\uparrow}$	**0.01200**$^{\Rightarrow\uparrow}$	**0.01260**$^{\Rightarrow\uparrow}$	**0.01253**$^{\Rightarrow\uparrow}$
	RPT	**0.04330**$^{\Rightarrow\uparrow\blacktriangle}$	**0.66295**$^{\Rightarrow\uparrow\blacktriangle}$	**0.47000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.40067**$^{\Rightarrow\uparrow\blacktriangle}$	**0.23640**$^{\Rightarrow\uparrow\blacktriangle}$	**0.29822**$^{\Rightarrow\uparrow\blacktriangle}$	**0.03127**$^{\Rightarrow\uparrow\blacktriangle}$	**0.50853**$^{\Rightarrow\uparrow\blacktriangle}$	**0.11000**$^{\Rightarrow\uparrow\blacktriangle}$	**0.12534**$^{\Rightarrow\uparrow\blacktriangle}$	**0.10260**$^{\Rightarrow\uparrow\blacktriangle}$	**0.08134**$^{\Rightarrow\uparrow\blacktriangle}$

B.2 Robustness in Varying of Training and Testing Boundaries

Table B.12: Retrieval effectiveness Boundaries ASIS Environment (75-25)%

ASIS*		ASISWOR test-bed						ASISWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.03400	0.45340	0.18400	0.14800	0.09960	0.06819	0.01935	0.31112	0.00400	0.01200	0.01440	0.01261
	CORI	0.03504	0.46613	0.18800	0.14666	0.09920	0.07337	0.02471	0.37178	0.00400	0.01200	0.01520	0.01492
	R/RP/RT	**0.03760**$^{\Rightarrow\Uparrow}$	**0.47962**$^{\Rightarrow\Uparrow}$	**0.18800**$^{\Uparrow}$	**0.16001**$^{\Rightarrow\Uparrow}$	**0.10680**	**0.07743**$^{\Rightarrow\Uparrow}$	**0.02317**$^{\Uparrow}$	**0.32345**$^{\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.01600**$^{\Rightarrow\Uparrow}$	**0.01442**$^{\Uparrow}$
	RPT	**0.05120**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.67614**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.42800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.37467**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22760**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29953**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03481**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49633**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09200**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11734**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09360**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09067**$^{\Rightarrow\Uparrow\blacktriangle}$
30%	Taily	0.03328	0.45050	0.18400	0.14533	0.09880	0.06666	0.01675	0.27231	0.00400	0.01200	0.01400	0.01139
	CORI	0.03516	0.46026	0.18800	0.14400	0.10000	0.07301	0.02143	0.33272	0.00400	0.01200	0.01400	0.01286
	R/RP/RT	**0.03764**$^{\Rightarrow\Uparrow}$	**0.48064**$^{\Rightarrow\Uparrow}$	**0.18800**$^{\uparrow}$	**0.16001**$^{\Rightarrow\Uparrow}$	**0.10680**	**0.07755**$^{\Rightarrow\Uparrow}$	**0.02317**$^{\Rightarrow\Uparrow}$	**0.32345**$^{\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.01600**$^{\Rightarrow\Uparrow}$	**0.01441**$^{\Rightarrow\Uparrow}$
	RPT	**0.05124**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.67717**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.42800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.37467**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22760**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29960**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03481**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49633**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11734**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09360**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09071**$^{\Rightarrow\Uparrow\blacktriangle}$
40%	Taily	0.03320	0.44805	0.18400	0.14267	0.09840	0.06604	0.01535	0.25725	0.00400	0.01200	0.01360	0.01082
	CORI	0.03504	0.46452	0.18800	0.14400	0.10040	0.07252	0.01887	0.30760	0.00400	0.01200	0.01360	0.01173
	R/RP/RT	**0.03764**$^{\Rightarrow\Uparrow}$	**0.48064**$^{\Rightarrow\Uparrow}$	**0.18800**$^{\uparrow}$	**0.16001**$^{\Rightarrow\Uparrow}$	**0.10680**	**0.07755**$^{\Rightarrow\Uparrow}$	**0.02317**$^{\Rightarrow\Uparrow}$	**0.32345**$^{\Rightarrow\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.01600**$^{\Rightarrow\Uparrow}$	**0.01442**$^{\Rightarrow\Uparrow}$
	RPT	**0.05124**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.67717**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.42800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.37467**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22760**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29960**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03481**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49633**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11734**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09360**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09071**$^{\Rightarrow\Uparrow\blacktriangle}$
50%	Taily	0.03320	0.45329	0.18800	0.14266	0.09880	0.06827	0.01403	0.25068	0.00400	0.01200	0.01360	0.01051
	CORI	0.03488	0.46234	0.18800	0.14400	0.09960	0.07186	0.01647	0.27638	0.00400	0.01200	0.01360	0.01093
	R/RP/RT	**0.03764**$^{\Rightarrow\Uparrow}$	**0.48064**$^{\Rightarrow\Uparrow}$	**0.18800**	**0.16001**$^{\Rightarrow\Uparrow}$	**0.10680**$^{\Rightarrow\Uparrow}$	**0.07755**$^{\Rightarrow\Uparrow}$	**0.02313**$^{\Rightarrow\Uparrow}$	**0.32334**$^{\Rightarrow\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01733**$^{\Rightarrow\Uparrow}$	**0.01600**$^{\Rightarrow\Uparrow}$	**0.01440**$^{\Rightarrow\Uparrow}$
	RPT	**0.05124**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.67717**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.42800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.37467**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.22760**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.29960**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.03481**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.49633**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.08800**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.11734**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09360**$^{\Rightarrow\Uparrow\blacktriangle}$	**0.09071**$^{\Rightarrow\Uparrow\blacktriangle}$

B.2 Robustness in Varying of Training and Testing Boundaries

Table B.13: Retrieval effectiveness Boundaries U Environment (25-75)%

U*		UWOR test-bed						UWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.02884	0.51029	0.20000	0.14933	0.10120	0.09652	0.02428	0.46379	0.00933	0.01422	0.02107	0.02193
	CORI	0.02993	0.53092	0.19867	0.14889	0.10347	0.10061	0.02700	0.49789	0.00933	0.01333	0.02040	0.02425
	R/RP/RT/RPT	**0.03747**$^{\Rightarrow\Uparrow}$	**0.55592**$^{\Rightarrow\Uparrow}$	**0.29467**$^{\Rightarrow\Uparrow}$	**0.21289**$^{\Rightarrow\Uparrow}$	**0.13453**$^{\Rightarrow\Uparrow}$	**0.14621**$^{\Rightarrow\Uparrow}$	**0.03267**$^{\Rightarrow\Uparrow}$	**0.49521**$^{\Rightarrow\Uparrow}$	**0.02267**$^{\Rightarrow\Uparrow}$	**0.02844**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.03269**$^{\Rightarrow\Uparrow}$
30%	Taily	0.02901	0.52584	0.18800	0.14533	0.10107	0.09595	0.02290	0.44923	0.00800	0.01289	0.01867	0.02006
	CORI	0.03017	0.54043	0.18267	0.14800	0.10240	0.10160	0.02444	0.48781	0.00800	0.01289	0.01627	0.02123
	R/RP/RT/RPT	**0.04036**$^{\Rightarrow\Uparrow}$	**0.63239**$^{\Rightarrow\Uparrow}$	**0.29867**$^{\Rightarrow\Uparrow}$	**0.22444**$^{\Rightarrow\Uparrow}$	**0.14347**$^{\Rightarrow\Uparrow}$	**0.15945**$^{\Rightarrow\Uparrow}$	**0.03304**$^{\Rightarrow\Uparrow}$	**0.50739**$^{\Rightarrow\Uparrow}$	**0.02133**$^{\Rightarrow\Uparrow}$	**0.02844**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.03363**$^{\Rightarrow\Uparrow}$
40%	Taily	0.02886	0.52253	0.18800	0.14178	0.09667	0.09439	0.02154	0.42333	0.00933	0.01244	0.01627	0.01849
	CORI	0.03021	0.53769	0.18800	0.14311	0.09920	0.09985	0.02240	0.44287	0.00800	0.01244	0.01520	0.01902
	R/RP/RT/RPT	**0.04063**$^{\Rightarrow\Uparrow}$	**0.63769**$^{\Rightarrow\Uparrow}$	**0.29867**$^{\Rightarrow\Uparrow}$	**0.22266**$^{\Rightarrow\Uparrow}$	**0.14373**$^{\Rightarrow\Uparrow}$	**0.16040**$^{\Rightarrow\Uparrow}$	**0.03304**$^{\Rightarrow\Uparrow}$	**0.50739**$^{\Rightarrow\Uparrow}$	**0.02133**$^{\Rightarrow\Uparrow}$	**0.02844**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.03362**$^{\Rightarrow\Uparrow}$
50%	Taily	0.02864	0.51968	0.19067	0.14311	0.09640	0.09426	0.02072	0.41641	0.00667	0.01244	0.01480	0.01755
	CORI	0.03004	0.53467	0.19467	0.14400	0.09907	0.10061	0.02065	0.41936	0.00800	0.01200	0.01453	0.01767
	R/RP/RT/RPT	**0.04063**$^{\Rightarrow\Uparrow}$	**0.63769**$^{\Rightarrow\Uparrow}$	**0.29867**$^{\Rightarrow\Uparrow}$	**0.22266**$^{\Rightarrow\Uparrow}$	**0.14373**$^{\Rightarrow\Uparrow}$	**0.16040**$^{\Rightarrow\Uparrow}$	**0.03304**$^{\Rightarrow\Uparrow}$	**0.50739**$^{\Rightarrow\Uparrow}$	**0.02133**$^{\Rightarrow\Uparrow}$	**0.02844**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.03362**$^{\Rightarrow\Uparrow}$

Table B.14: Retrieval effectiveness Boundaries U Environment (50-50)%

U*		UWOR test-bed						UWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.02987	0.48763	0.21400	0.16266	0.10640	0.09912	0.02601	0.44331	0.00200	0.01066	0.01400	0.01817
	CORI	0.03121	0.51306	0.21200	0.16067	0.10900	0.10192	0.02845	0.46244	0.00400	0.01066	0.01580	0.02147
	R/RP/RT/RPT	**0.03719**$^{\Rightarrow\Uparrow}$	**0.53590**$^{\Rightarrow\Uparrow}$	**0.31400**$^{\Rightarrow\Uparrow}$	**0.23000**$^{\Rightarrow\Uparrow}$	**0.14440**$^{\Rightarrow\Uparrow}$	**0.14827**$^{\Rightarrow\Uparrow}$	**0.03367**$^{\Rightarrow\Uparrow}$	**0.48645**$^{\Rightarrow\Uparrow}$	**0.01600**$^{\Rightarrow\Uparrow}$	**0.01933**$^{\Rightarrow\Uparrow}$	**0.02020**$^{\Rightarrow\Uparrow}$	**0.02717**$^{\Rightarrow\Uparrow}$
30%	Taily	0.03013	0.50384	0.19400	0.15534	0.10740	0.09799	0.02429	0.41880	0.00200	0.00933	0.01180	0.01597
	CORI	0.03167	0.52371	0.19200	0.16067	0.10840	0.10633	0.02621	0.45571	0.00200	0.01000	0.01100	0.01770
	R/RP/RT/RPT	**0.04074**$^{\Rightarrow\Uparrow}$	**0.61001**$^{\Rightarrow\Uparrow}$	**0.32600**$^{\Rightarrow\Uparrow}$	**0.24066**$^{\Rightarrow\Uparrow}$	**0.15300**$^{\Rightarrow\Uparrow}$	**0.16241**$^{\Rightarrow\Uparrow}$	**0.03447**$^{\Rightarrow\Uparrow}$	**0.49867**$^{\Rightarrow\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01933**$^{\Rightarrow\Uparrow}$	**0.02080**$^{\Rightarrow\Uparrow}$	**0.02829**$^{\Rightarrow\Uparrow}$
40%	Taily	0.03025	0.50548	0.19400	0.15267	0.10320	0.09691	0.02273	0.40062	0.00200	0.00933	0.01060	0.01442
	CORI	0.03191	0.52669	0.19600	0.15467	0.10560	0.10419	0.02365	0.42113	0.00200	0.00933	0.00980	0.01511
	R/RP/RT/RPT	**0.04105**$^{\Rightarrow\Uparrow}$	**0.61602**$^{\Rightarrow\Uparrow}$	**0.32600**$^{\Rightarrow\Uparrow}$	**0.23933**$^{\Rightarrow\Uparrow}$	**0.15280**$^{\Rightarrow\Uparrow}$	**0.16327**$^{\Rightarrow\Uparrow}$	**0.03447**$^{\Rightarrow\Uparrow}$	**0.49867**$^{\Rightarrow\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01933**$^{\Rightarrow\Uparrow}$	**0.02080**$^{\Rightarrow\Uparrow}$	**0.02828**$^{\Rightarrow\Uparrow}$
50%	Taily	0.02993	0.49919	0.19800	0.15467	0.10260	0.09591	0.02165	0.38998	0.00200	0.00933	0.00900	0.01342
	CORI	0.03183	0.52362	0.20400	0.15600	0.10560	0.10496	0.02153	0.39523	0.00200	0.00867	0.00900	0.01354
	R/RP/RT/RPT	**0.04105**$^{\Rightarrow\Uparrow}$	**0.61602**$^{\Rightarrow\Uparrow}$	**0.32600**$^{\Rightarrow\Uparrow}$	**0.23933**$^{\Rightarrow\Uparrow}$	**0.15280**$^{\Rightarrow\Uparrow}$	**0.16327**$^{\Rightarrow\Uparrow}$	**0.03447**$^{\Rightarrow\Uparrow}$	**0.49867**$^{\Rightarrow\Uparrow}$	**0.01200**$^{\Rightarrow\Uparrow}$	**0.01933**$^{\Rightarrow\Uparrow}$	**0.02080**$^{\Rightarrow\Uparrow}$	**0.02828**$^{\Rightarrow\Uparrow}$

Table B.15: Retrieval effectiveness Boundaries U Environment (75-25)%

U*		UWOR test-bed						UWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	Taily	0.03507	0.47482	0.22400	0.17866	0.11840	0.11278	0.03019	0.43681	0.00400	0.01600	0.01560	0.01998
	CORI	0.03763	0.51966	0.21200	0.17600	0.12240	0.11746	0.03335	0.44762	0.00400	0.01600	0.01680	0.02356
	R/RP/RT/RPT	**0.04650**$^{\Rightarrow\Uparrow}$	**0.56646**$^{\Rightarrow\Uparrow}$	**0.30400**$^{\Rightarrow\Uparrow}$	**0.24400**$^{\Rightarrow\Uparrow}$	**0.16560**$^{\Rightarrow\Uparrow}$	**0.16691**$^{\Rightarrow\Uparrow}$	**0.04098**$^{\Rightarrow\Uparrow}$	**0.47438**$^{\Rightarrow\Uparrow}$	**0.02000**$^{\Rightarrow\Uparrow}$	**0.02666**$^{\Rightarrow\Uparrow}$	**0.01880**$^{\Rightarrow\Uparrow}$	**0.02961**$^{\Rightarrow\Uparrow}$
30%	Taily	0.03535	0.50606	0.20400	0.16800	0.12040	0.11226	0.02739	0.40966	0.00400	0.01600	0.01440	0.01751
	CORI	0.03787	0.53070	0.20400	0.17466	0.12160	0.12608	0.03007	0.44056	0.00400	0.01600	0.01440	0.01964
	R/RP/RT/RPT	**0.05024**$^{\Rightarrow\Uparrow}$	**0.61879**$^{\Rightarrow\Uparrow}$	**0.31200**$^{\Rightarrow\Uparrow}$	**0.25600**$^{\Rightarrow\Uparrow}$	**0.17280**$^{\Rightarrow\Uparrow}$	**0.18182**$^{\Rightarrow\Uparrow}$	**0.04202**$^{\Rightarrow\Uparrow}$	**0.49392**$^{\Rightarrow\Uparrow}$	**0.02000**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.01960**$^{\Rightarrow\Uparrow}$	**0.03160**$^{\Rightarrow\Uparrow}$
40%	Taily	0.03603	0.51588	0.20400	0.16800	0.11720	0.11244	0.02499	0.38674	0.00400	0.01600	0.01360	0.01580
	CORI	0.03835	0.53527	0.20400	0.17067	0.11960	0.12485	0.02619	0.41032	0.00400	0.01600	0.01320	0.01642
	R/RP/RT/RPT	**0.05065**$^{\Rightarrow\Uparrow}$	**0.62859**$^{\Rightarrow\Uparrow}$	**0.31600**$^{\Rightarrow\Uparrow}$	**0.25600**$^{\Rightarrow\Uparrow}$	**0.17280**$^{\Rightarrow\Uparrow}$	**0.18384**$^{\Rightarrow\Uparrow}$	**0.04202**$^{\Rightarrow\Uparrow}$	**0.49392**$^{\Rightarrow\Uparrow}$	**0.02000**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.01960**$^{\Rightarrow\Uparrow}$	**0.03159**$^{\Rightarrow\Uparrow}$
50%	Taily	0.03555	0.50544	0.19600	0.16666	0.11600	0.11002	0.02335	0.36966	0.00400	0.01600	0.01280	0.01498
	CORI	0.03823	0.53164	0.20400	0.16933	0.11840	0.12526	0.02279	0.37280	0.00400	0.01600	0.01280	0.01461
	R/RP/RT/RPT	**0.05065**$^{\Rightarrow\Uparrow}$	**0.62859**$^{\Rightarrow\Uparrow}$	**0.31600**$^{\Rightarrow\Uparrow}$	**0.25600**$^{\Rightarrow\Uparrow}$	**0.17280**$^{\Rightarrow\Uparrow}$	**0.18384**$^{\Rightarrow\Uparrow}$	**0.04202**$^{\Rightarrow\Uparrow}$	**0.49392**$^{\Rightarrow\Uparrow}$	**0.02000**$^{\Rightarrow\Uparrow}$	**0.02800**$^{\Rightarrow\Uparrow}$	**0.01960**$^{\Rightarrow\Uparrow}$	**0.03159**$^{\Rightarrow\Uparrow}$

B.3 Reputation-based and CORI Approaches

Table B.16: Retrieval effectiveness (α R + (1-α) CORI) DL Environment

DL*		DLWOR test-bed						DLWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	RT	0.02828	0.50166	0.11800	0.12267	0.08670	0.07433	0.02575	0.48813	**0.02700**	**0.02466**	0.02770	0.02721
	RPT	0.04312	0.71150	0.28400	0.28500	0.20260	0.23551	0.03610	**0.63776**	0.17700	0.17967	0.14460	0.14692
	RCORI	**0.03005**$^{\triangle}$	**0.53005**	**0.16800**$^{\blacktriangle}$	**0.14233**$^{\blacktriangle}$	**0.09420**$^{\blacktriangle}$	**0.08992**$^{\blacktriangle}$	**0.02684**	**0.49555**	0.02400	0.02433	**0.02920**	**0.02939**
	RPCORI	**0.16208**$^{\blacktriangle}$	**0.72663**	**0.35600**$^{\blacktriangle}$	**0.31900**$^{\blacktriangle}$	**0.20920**	**0.26443**$^{\blacktriangle}$	**0.16193**$^{\blacktriangle}$	0.56093$^{\blacktriangledown}$	**0.22500**$^{\blacktriangle}$	**0.24934**$^{\blacktriangle}$	**0.17440**$^{\blacktriangle}$	**0.17138**$^{\triangle}$
30%	RT	0.02968	0.54172	0.12200	0.12600	0.09010	0.07990	0.02635	0.50571	**0.02600**	**0.02333**	0.02690	0.02776
	RPT	0.04645	0.78327	0.28600	0.29033	**0.21480**	0.26318	0.03831	**0.68640**	0.16700	0.17367	0.14390	0.15043
	RCORI	**0.03032**	**0.55983**	**0.16000**$^{\blacktriangle}$	**0.13733**$^{\blacktriangle}$	**0.09460**$^{\triangle}$	**0.09366**$^{\blacktriangle}$	**0.02715**	**0.51673**	0.02400	0.02166	**0.02750**	**0.02860**
	RPCORI	**0.15488**$^{\blacktriangle}$	**0.80144**	**0.31900**	**0.30333**	0.20790	**0.26804**	**0.15420**$^{\blacktriangle}$	0.61701$^{\blacktriangledown}$	**0.19700**$^{\triangle}$	**0.23567**$^{\blacktriangle}$	**0.17420**$^{\blacktriangle}$	**0.17192**
40%	RT	0.02993	0.55529	0.11600	0.12500	0.09060	0.07992	0.02615	0.51061	**0.02400**	**0.02266**	**0.02640**	0.02763
	RPT	0.04783	0.81716	0.28600	0.29066	**0.22000**	**0.27281**	0.03905	**0.70600**	0.16200	0.16900	0.14190	0.15079
	RCORI	**0.03065**$^{\triangle}$	**0.57668**$^{\blacktriangle}$	**0.15700**$^{\blacktriangle}$	**0.13200**$^{\triangle}$	**0.09260**	**0.09049**$^{\blacktriangle}$	**0.02683**	**0.51610**	**0.02400**	0.02100	0.02580	**0.02799**
	RPCORI	**0.15016**$^{\blacktriangle}$	**0.83224**	**0.29900**	**0.29033**	0.20650$^{\triangledown}$	0.25706	**0.14994**$^{\blacktriangle}$	0.64429$^{\blacktriangledown}$	**0.19000**$^{\triangle}$	**0.22167**$^{\blacktriangle}$	**0.17020**$^{\blacktriangle}$	**0.17436**$^{\triangle}$
50%	RT	0.03024	0.57382	0.11700	0.12333	0.09040	0.08138	0.02589	0.50896	0.02300	**0.02067**	0.02440	0.02682
	RPT	0.04868	**0.84544**	**0.28600**	**0.29100**	**0.22130**	**0.27728**	0.03910	**0.71020**	0.15900	0.16433	0.13800	0.14797
	RCORI	**0.03058**	**0.57495**	**0.14200**$^{\blacktriangle}$	**0.12533**	**0.09100**	**0.08540**$^{\blacktriangle}$	**0.02663**$^{\triangle}$	**0.52481**$^{\triangle}$	**0.02400**	**0.02067**	**0.02510**$^{\triangle}$	**0.02765**$^{\blacktriangle}$
	RPCORI	**0.14819**$^{\blacktriangle}$	0.83419	0.26600	0.27299$^{\triangledown}$	0.20070$^{\triangledown}$	0.24422$^{\triangledown}$	**0.14736**$^{\blacktriangle}$	0.66072$^{\blacktriangledown}$	**0.18400**$^{\triangle}$	**0.21000**$^{\blacktriangle}$	**0.16770**$^{\blacktriangle}$	**0.17079**$^{\triangle}$

Table B.17: Retrieval effectiveness (α R + (1-α) CORI) ASIS Environment

ASIS*		ASISWOR test-bed						ASISWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	RT	0.02907	0.48191	0.17000	0.12900	**0.08420**	0.07851	0.02907	0.48191	0.17000	0.12900	**0.08420**	0.07851
	RPT	0.04619	**0.75096**	**0.27300**	**0.26567**	**0.20100**	**0.22530**	0.04619	**0.75096**	**0.27300**	**0.26567**	**0.20100**	**0.22530**
	RCORI	**0.02943**	**0.48787**$^{\triangle}$	**0.17200**	**0.13000**	0.08390	**0.08034**	**0.02943**$^{\triangle}$	**0.48787**$^{\blacktriangle}$	**0.17200**	**0.13000**	0.08390$^{\blacktriangledown}$	**0.08034**$^{\blacktriangle}$
	RPCORI	**0.14633**$^{\blacktriangle}$	0.74551	0.26300	0.26234	0.18390$^{\blacktriangledown}$	0.20358$^{\triangledown}$	**0.14633**$^{\blacktriangle}$	0.74551$^{\blacktriangledown}$	0.26300$^{\blacktriangledown}$	0.26234$^{\blacktriangledown}$	0.18390$^{\blacktriangledown}$	0.20358$^{\blacktriangledown}$
30%	RT	0.02908	0.48217	0.17000	0.12900	0.08420	0.07853	0.02908	0.48217	0.17000	0.12900	0.08420	0.07853
	RPT	0.04621	**0.75138**	**0.27300**	**0.26567**	**0.20100**	**0.22534**	0.04621	**0.75138**	**0.27300**	**0.26567**	**0.20100**	**0.22534**
	RCORI	**0.02939**$^{\blacktriangle}$	**0.48552**$^{\triangle}$	**0.17200**	**0.13033**	**0.08490**	**0.07974**	**0.02939**	**0.48552**$^{\triangle}$	**0.17200**	**0.13033**	**0.08490**	**0.07974**$^{\blacktriangle}$
	RPCORI	**0.14426**$^{\blacktriangle}$	0.75113	0.24600$^{\blacktriangledown}$	0.25000$^{\triangledown}$	0.18040$^{\blacktriangledown}$	0.19600$^{\blacktriangledown}$	**0.14426**$^{\blacktriangle}$	0.75113$^{\blacktriangledown}$	0.24600$^{\blacktriangledown}$	0.25000$^{\blacktriangledown}$	0.18040$^{\blacktriangledown}$	0.19600$^{\blacktriangledown}$
40%	RT	0.02908	0.48217	**0.17000**	**0.12900**	0.08420	0.07853	0.02908	0.48217	**0.17000**	**0.12900**	0.08420	0.07853
	RPT	0.04621	**0.75138**	**0.27300**	**0.26567**	**0.20100**	**0.22534**	0.04621	**0.75138**	**0.27300**	**0.26567**	**0.20100**	**0.22534**
	RCORI	**0.02924**	**0.48326**	0.16900	0.12867	**0.08440**	**0.07889**	**0.02924**$^{\triangle}$	**0.48326**$^{\triangle}$	0.16900	0.12867	**0.08440**	**0.07889**$^{\blacktriangle}$
	RPCORI	**0.14348**$^{\blacktriangle}$	0.75110	0.23800$^{\blacktriangledown}$	0.24167$^{\blacktriangledown}$	0.17780$^{\blacktriangledown}$	0.19163$^{\blacktriangledown}$	**0.14348**$^{\blacktriangle}$	0.75110$^{\blacktriangledown}$	0.23800$^{\blacktriangledown}$	0.24167$^{\blacktriangledown}$	0.17780$^{\blacktriangledown}$	0.19163$^{\blacktriangledown}$
50%	RT	0.02907	0.48210	**0.17000**	**0.12900**	0.08420	0.07853	0.02907	0.48210	**0.17000**	**0.12900**	0.08420	0.07853
	RPT	0.04621	**0.75138**	**0.27300**	**0.26567**	**0.20100**	**0.22534**	0.04621	**0.75138**	**0.27300**	**0.26567**	**0.20100**	**0.22534**
	RCORI	**0.02920**	**0.48299**	**0.17000**	**0.12900**	**0.08430**	**0.07885**	**0.02920**	**0.48299**	**0.17000**	**0.12900**	**0.08430**	**0.07885**$^{\triangle}$
	RPCORI	**0.14306**$^{\blacktriangle}$	0.74736$^{\triangledown}$	0.23200$^{\blacktriangledown}$	0.23767$^{\blacktriangledown}$	0.17640$^{\blacktriangledown}$	0.18973$^{\blacktriangledown}$	**0.14306**$^{\blacktriangle}$	0.74736$^{\blacktriangledown}$	0.23200$^{\blacktriangledown}$	0.23767$^{\blacktriangledown}$	0.17640$^{\blacktriangledown}$	0.18973$^{\blacktriangledown}$

Table B.18: Retrieval effectiveness (α R + (1-α) CORI) U Environment

U*		UWOR test-bed						UWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
20%	RT	0.03987	0.56138	0.29600	0.21233	0.13380	0.16051	0.03467	0.48008	0.02900	0.03700	0.03700	0.03696
	RPT	0.04734	0.67614	0.23700	0.22766	0.17470	0.17353	0.03772	**0.53016**	0.10600	0.09733	0.08000	0.07205
	RCORI	**0.04146**	**0.61666**$^{\blacktriangle}$	**0.30500**	**0.22299**$^{\triangle}$	**0.14190**$^{\triangle}$	**0.16815**	**0.03586**	**0.49901**	**0.03200**	**0.04500**$^{\blacktriangle}$	**0.04770**$^{\blacktriangle}$	**0.04350**$^{\blacktriangle}$
	RPCORI	**0.17005**$^{\blacktriangle}$	**0.71491**	**0.37400**$^{\blacktriangle}$	**0.32834**$^{\blacktriangle}$	**0.21930**$^{\blacktriangle}$	**0.25903**$^{\blacktriangle}$	**0.16589**$^{\blacktriangle}$	0.47932$^{\blacktriangledown}$	**0.19900**$^{\blacktriangle}$	**0.21500**$^{\blacktriangle}$	**0.17160**$^{\blacktriangle}$	**0.14526**$^{\blacktriangle}$
30%	RT	0.04140	0.61208	0.29100	0.21500	0.13800	0.16450	0.03501	0.50751	0.02700	0.03666	0.03790	0.03813
	RPT	0.05020	0.74535	0.23800	0.23000	0.18070	0.18808	0.03800	**0.55371**	0.10600	0.09700	0.08010	0.07295
	RCORI	**0.04213**	**0.62307**	**0.30200**	**0.22167**	**0.14140**$^{\triangle}$	**0.17143**$^{\triangle}$	**0.03659**$^{\triangle}$	**0.53137**	**0.02900**	**0.03900**$^{\blacktriangle}$	**0.04290**$^{\blacktriangle}$	**0.04234**$^{\blacktriangle}$
	RPCORI	**0.15723**$^{\blacktriangle}$	**0.76668**	**0.35200**$^{\blacktriangle}$	**0.30534**$^{\blacktriangle}$	**0.21700**$^{\blacktriangle}$	**0.25011**$^{\blacktriangle}$	**0.15477**$^{\blacktriangle}$	0.54009	**0.18300**$^{\blacktriangle}$	**0.20000**$^{\blacktriangle}$	**0.16910**$^{\blacktriangle}$	**0.14437**$^{\blacktriangle}$
40%	RT	0.04151	0.62181	**0.29200**	0.21466	0.13800	0.16511	0.03501	0.50751	**0.02700**	0.03666	0.03790	0.03812
	RPT	0.05036	0.75520	0.23800	0.23000	0.18120	0.18891	0.03800	0.55371	0.10600	0.09700	0.08010	0.07295
	RCORI	**0.04220**	**0.62343**	0.29100	**0.21833**	**0.14060**$^{\triangle}$	**0.16903**$^{\triangle}$	**0.03657**$^{\triangle}$	**0.53168**	**0.02700**	**0.03766**	**0.04010**$^{\blacktriangle}$	**0.04135**$^{\blacktriangle}$
	RPCORI	**0.15149**$^{\blacktriangle}$	**0.79214**$^{\triangle}$	**0.32700**$^{\blacktriangle}$	**0.29134**$^{\blacktriangle}$	**0.21040**$^{\blacktriangle}$	**0.23829**$^{\blacktriangle}$	**0.14792**$^{\blacktriangle}$	**0.56387**	**0.17900**$^{\blacktriangle}$	**0.19133**$^{\blacktriangle}$	**0.16450**$^{\blacktriangle}$	**0.14072**$^{\blacktriangle}$
50%	RT	0.04151	0.62181	0.29200	0.21466	0.13800	0.16511	0.03501	0.50751	**0.02700**	0.03666	0.03790	0.03812
	RPT	0.05036	0.75520	0.23800	0.23000	0.18120	0.18891	0.03800	0.55371	0.10600	0.09700	0.08010	0.07295
	RCORI	**0.04236**$^{\triangle}$	**0.62615**	**0.29300**	**0.21833**	**0.14080**$^{\blacktriangle}$	**0.16815**$^{\triangle}$	**0.03641**$^{\blacktriangle}$	**0.53691**$^{\blacktriangle}$	**0.02700**	**0.03700**	**0.03890**$^{\blacktriangle}$	**0.04049**$^{\blacktriangle}$
	RPCORI	**0.14791**$^{\blacktriangle}$	**0.80259**$^{\blacktriangle}$	**0.31200**$^{\blacktriangle}$	**0.27767**$^{\blacktriangle}$	**0.20420**$^{\blacktriangle}$	**0.22575**$^{\blacktriangle}$	**0.14355**$^{\blacktriangle}$	**0.58249**	**0.16700**$^{\blacktriangle}$	**0.17900**$^{\blacktriangle}$	**0.15940**$^{\blacktriangle}$	**0.13669**$^{\blacktriangle}$

B.4 Reputation-based Approaches Under Noisy Information

Table B.19: Reputation-based Retrieval Effectiveness on Noisy Data

DL*		DLWOR test-bed						DLWR test-bed					
%	**Method**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**	**Precision**	**Recall**	**P@10**	**P@30**	**P@100**	**MAP**
2%	RN	0.02543	0.44928	**0.16040**	**0.12593**	0.08086	**0.08032**	0.02334	0.41971	**0.02660**	**0.02480**	**0.02762**	0.02429
	RTN	0.02635	0.45029	**0.15100**	**0.12480**	0.08164	**0.07909**	0.02394	0.41998	**0.02620**	**0.02493**	**0.02768**	0.02467
3%	RN	0.02609	0.46062	**0.15880**	**0.12633**	0.08246	**0.08386**	0.02343	0.43138	**0.02680**	0.02326	**0.02764**	0.02429
	RTN	0.02652	0.45445	**0.15140**	**0.12540**	0.08254	**0.07938**	0.02413	0.42479	**0.02700**	**0.02500**	**0.02788**	0.02491
4%	RN	0.02665	0.46138	**0.15280**	**0.12546**	0.08318	**0.08064**	0.02372	0.44107	0.02500	**0.02386**	**0.02790**	0.02458
	RTN	0.02665	0.46138	**0.15280**	**0.12546**	0.08318	**0.08064**	0.02424	0.42517	**0.02720**	**0.02500**	**0.02792**	0.02492
5%	RN	0.02633	0.47255	**0.16500**	**0.13040**	0.08444	**0.08590**	0.02390	0.44277	0.02540	**0.02346**	**0.02764**	0.02474
	RTN	0.02668	0.46166	**0.15360**	**0.12646**	0.08326	**0.08100**	0.02428	0.42701	**0.02800**	**0.02493**	**0.02810**	0.02509
ASIS*		ASISWOR test-bed						ASISWR test-bed					
2%	RN	0.02880	0.47671	0.17040	0.12860	0.08366	0.07785	0.02378	0.42203	0.02200	0.02706	0.02508	0.02303
	RTN	0.02870	0.47437	0.17040	0.12800	0.08316	0.07780	0.02223	0.39428	0.02220	0.02680	0.02452	0.02229
3%	RN	0.02879	0.47627	0.16980	0.12827	0.08368	0.07770	0.02380	0.42238	0.02200	0.02706	0.02498	0.02305
	RTN	0.02873	0.47499	0.17020	0.12807	0.08330	0.07791	0.02224	0.39426	0.02220	0.02680	0.02454	0.02230
4%	RN	0.02879	0.47680	0.16980	0.12807	0.08366	0.07779	0.02379	0.42218	0.02200	0.02700	0.02488	0.02294
	RTN	0.02876	0.47548	0.17020	0.12800	0.08334	0.07795	0.02225	0.39459	0.02220	0.02680	0.02454	0.02230
5%	RN	0.02880	0.47575	0.16960	0.12827	0.08368	0.07788	0.02379	0.42214	0.02200	0.02713	0.02490	0.02293
	RTN	0.02876	0.47549	0.17020	0.12813	0.08336	0.07797	0.02225	0.39468	0.02220	0.02680	0.02452	0.02230
U*		UWOR test-bed						UWR test-bed					
2%	RN	0.03878	0.55090	0.28120	0.20586	0.12874	0.14628	0.03358	**0.47003**	**0.02840**	0.03626	0.03610	0.03496
	RTN	0.03928	**0.56682**	0.28360	**0.20806**	0.13030	0.15042	**0.03422**	**0.48695**	**0.03000**	**0.03960**	**0.04008**	**0.03783**
3%	RN	0.03851	0.54957	0.27820	0.20426	0.12792	0.14512	**0.03422**	**0.48515**	0.02720	**0.03680**	**0.03688**	**0.03617**
	RTN	0.03930	**0.56769**	0.28340	**0.20800**	0.13020	0.15030	**0.03419**	**0.48712**	**0.02980**	**0.03940**	**0.04020**	**0.03775**
4%	RN	0.03835	0.54662	0.27640	0.20486	0.12740	0.14461	0.03351	**0.46980**	**0.02780**	**0.03646**	0.03606	0.03487
	RTN	0.03927	**0.56721**	0.28260	0.20753	0.13004	0.15010	**0.03418**	**0.48712**	**0.02980**	**0.03933**	**0.04016**	**0.03773**
5%	RN	0.03827	0.54831	0.27640	0.20293	0.12712	0.14572	0.03348	0.46821	**0.02800**	0.03613	0.03606	0.03467
	RTN	0.03926	**0.56647**	0.28300	**0.20813**	0.13008	0.15017	**0.03413**	**0.48599**	**0.02960**	**0.03933**	**0.04006**	**0.03762**